Integrating speech with visual text in ~~the~~ multimodal interfaces

Yael Shmueli

Dissertation

Submitted in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy in Cognitive Science

in the Department of Computer Science

University College London

May 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

This work systematically investigates when and how combining speech output and visual text may facilitate processing and comprehension of sentences. It is proposed that a redundant multimodal presentation of speech and text has the potential for improving sentence processing but also for severely disrupting it. The effectiveness of the presentation is assumed to depend on the linguistic complexity of the sentence, the memory demands incurred by the selected multimodal configuration and the characteristics of the user.

The thesis employs both theoretical and empirical methods to examine this claim. At the theoretical front, the research makes explicit features of multimodal sentence presentation and of structures and processes involved in multimodal language processing. Two entities are presented: a multimodal design space (MMDS) and a multimodal user model (MMUM). The dimensions of the MMDS include aspects of (i) the sentence (linguistic complexity, c.f., Gibson, 1991), (ii) the presentation (configurations of media), and (iii) user cost (a function of the first two dimensions). The second entity, the MMUM, is a cognitive model of the user. The MMUM attempts to characterise the cognitive structures and processes underlying multimodal language processing, including the supervisory attentional mechanisms that coordinate the processing of language in parallel modalities. The model includes an account of individual differences in verbal working memory (WM) capacity (c.f., Just & Carpenter, 1992) and can predict the variation in the cognitive cost experienced by the user when presented with different contents in a variety of multimodal configurations.

The work attempts to validate through 3 controlled studies with users the central propositions of the MMUM. The experimental findings indicate the validity of some features of the MMUM but also the need for further refinement. Overall, they suggest that a durable text may reduce the processing cost of demanding sentences delivered by speech, whereas adding speech to such sentences when presented visually increases processing cost. Speech can be added to various visual forms of text only if the linguistic complexity of the sentence imposes a low to moderate load on the user. These conclusions are translated to a set of guidelines for effective multimodal presentation of sentences. A final study then examines the validity of some of these guidelines in an applied setting. Results highlight the need for an enhanced experimental control. However, they also demonstrate that the approach used in this research can validate specific assumptions regarding the relationship between cognitive cost, sentence complexity and multimodal configuration aspects and thereby to inform the design process of effective multimodal user interfaces.

Acknowledgments

This research would not have been possible without the support of many people, a few of whom I would like to thank here. I have benefited from the aid of two grants: a research studentship from the School of Informatics at City University and a Computer Science Department bursary from UCL. These have supported me while I completed my PhD.

Thanks are due first to my supervisor, Dr. John Dowell, for his guidance throughout this research and for the freedom he gave me to pursue my ideas while keeping me on the right track. I would also like to thank him for his patience, support, encouragement and friendship. It's been a long process. Thanks are also due to Dr. Nilli Lavie for guiding the design of experiment 1, to Dr. Alastair McClelland, for his help on the statistics and to Dr. David Green, for his advice in the design of the applied study.

Additional thanks go to my friends. To Ran Sariel, for adding useful features to the multimodal user interface in the applied study. To Anjy Pavell, for lending her voice to the applied study and also for being so kind as to read the whole manuscript thoroughly, providing useful comments. To Naama Friedman, for all her help and to Lilit, Ofra, Eyal, Itai, Eran and Abbey, for their continuous moral support.

Special thanks to the Shmueli family. To my mother, Miriam Shmueli, who died before the work was complete and to my father, Uri. Thanks for giving me a loving home and for being a steady source of support and encouragement. To my brother Rami, for the useful discussions and helpful criticism and to his wife Roni. I would also like to thank the members of the Friedland family who have endured this long journey with me: to Jeffrey, Techia, Oded, Butsi, Yael, Lior, Eran and Keren.

I am particularly grateful to my husband Omri for his invaluable help: for lending his voice to experiments 1-3, for copy editing this thesis and for our scientific discussions. Without his love and support, I would never have finished it. Finally, I would like to thank our 3 years old daughter Noa and baby Michael, being there puts everything into the right perspective.

Contents

List of Tables

List of figures

Chapter 1

Introduction

This chapter first describes the research problem investigated in this thesis: when and how speech can be combined with visual text to facilitate the user's processing and comprehension of content. The definitions used in this work and the benefits of multimodal systems are summarised prior to the description of current applications, domains of application and the state of the art in multimodal systems. Following a short review of usability research of multimodality, the chapter examines the adequacy of current human-computer-interaction (HCI) theories in addressing the design problem, considering the available guidelines, multimodal taxonomies and models and also three cognitive theories/architectures. Suggestions for an alternative framework of research, its adequacy and potential contribution follow. The structure of the thesis is outlined last.

1.1 The research problem: the effect of linguistic complexity on the integration of verbal material in various multimodal configurations

Rapid advances in speech technology suggest that interaction by speech will be a common feature of working with computers in this decade. It is assumed that speech will mostly be added to, rather than replace, the established visual media for presenting verbal information in various applications (e.g., in hand held devices, interactive learning systems, information kiosks and the world wide web (WWW)). In other applications, visual media may be added to the primary speech channel (e.g., in public address (PA) systems and telephone-based services). The use of speech output with visual displays of text raises novel cognitive design issues about when and how the two media should be integrated. Other work has tried to answer questions about the use of speech as an alternative to text (e.g., Wright, 2001). By contrast, this work is asking the questions about when and how speech can be combined with visual text to facilitate the user's processing and comprehension of content.

The questions are raised on the backdrop of a consensus in the literature that attempting to listen to speech while reading will always and inevitably lead to interference and an increased processing cost (e.g., Mayes, 1994). However, the evidence on which this consensus is based relates to the processing of different verbal materials where background speech is found to interfere with a concurrent reading task (Martin, Vogalter & Forlano, 1988; Wickens, 1992). In contrast, the focus of this work is on optimising the processing cost of redundant speech and text. The potential for visual text processing to be enhanced by concurrent listening and for speech processing to be enhanced by a simultaneous presentation of visual text is indicated by a small number of cross-modal priming (CMP) studies showing such effects at the individual word level (e.g., Greenwald, 1970; Lewis, 1972; Kirsner & Smith, 1974; Hanson, 1981). In addition, intuitively good use is made of

combinations of visual text and speech in many crafted presentations, such as television advertisements and pop music video clips. It is the aim of this work to discover when presenting sentences by means of redundant speech and visual text will reduce processing cost, as has been defined for the single word level. Understanding redundant multimodality would then form the basis for future studies of other forms of multimodality that will not be assessed by this work

The central factors to affect processing cost are assumed to be both sentence complexity and its representational form. The thesis proposes that a redundant multimodal presentation of speech and text may have the potential for reducing user processing cost, depending on the linguistic complexity of the sentence and the memory demands incurred by the presentation type. This proposal refers both to the use of text in addition to speech and to the use of speech in addition to text, although the change in processing cost is assumed to differ slightly in each case. Whereas a durable text may reduce the processing cost of spoken sentences under medium linguistic complexity conditions (e.g., for long-simple sentences), the addition of speech may reduce processing cost only when the linguistic complexity value of the sentence imposes a low cognitive load on the user. Under high cognitive load conditions, the added speech might disrupt the processing of visual text. The human factor too must be addressed: users differ in their processing capacity, which might affect their ability to use multimodal systems effectively. A thorough investigation of these factors is expected to inform the design of effective multimodal systems, supporting all users in their work.

1.2 Application issues: the emergence of multimodal systems

1.2.1 Definition of multimodal presentation

'Multimodal interfaces' has recently become a buzzword among HCI researchers, but there is a lack of consistency regarding its meaning probably due to the interdisciplinary nature of the field (Benoit, Martin, Pelachaud, Schomaker & Suhm, 1998). The next section describes the most important terms, their uses in the field and the definitions that appear in this work.

There has been discussion in the field about the meaning of and distinction between multimodal and multimedia applications and the related terminology. Several expressions are used: medium, media, mode, modality, multimedia and multimodal.
Psychologists use the term 'modality' explicitly in the context of the human senses (sight, hearing, touch, smell, taste, and balance). On the other hand, researchers in the computer science discipline often use the term 'media' for physical devices and the term 'modality' for "a way to use a media" (e.g., Coutaz, 1992). For example, the visual display media can use several modalities such as text, graphics, picture and video to present output to the user. For some, media is related to computer systems while modalities are related to human beings. In Maybury (1993), media refers both to the material object (e.g., paper, video) as well as the means by which information is conveyed (a sheet of paper with text on it), while modalities refer to the human senses employed to process incoming information (e.g., vision, audition). Other researchers identify multimodality (multiple modalities)

with user input (from user to system) and multimedia-lity (multiple media) with system outputs (from system to user) (e.g., Oviatt, 1999, Oviatt & Cohen, 2000, W3C organisation: Multimodal Requirements for Voice Mark-up Languages). Finally, some researchers claim that the difference between multimedia and multimodal systems is in the ability of multimodal systems to model the content of the presented information at a high level of abstraction (e.g., Vernier & Nigay, 2000, Benoit et al., 1998). Clearly, there is no overall agreement in the field on terminology.

In this work, the definition used for the term 'multimodality' relies on the traditional terminology defined by Dix (1993). According to Dix, multimedia systems use different media to communicate supplementary, additional or redundant information. This may take the form of multiple sensory channels (e.g., multimedia multimodal systems), but it may also take the form of different types of visual output (e.g., textual, graphical and iconic). On the other hand, multimodal systems have been developed to take advantage of the multi sensory nature of humans as information processors. Utilising more than one human sensory channel, or mode of communication, these systems make much fuller use of the auditory channel and to a lesser extent, the tactile channel, to improve the interactive nature of the system. Thus, multimodal systems increase the bandwidth of HCI. This definition of multimodality does not exclude the use of multiple input devices in such systems, although the investigation of multimodal input is beyond the scope of this thesis. The next section describes the potential benefits of an audio-visual multimodal presentation of verbal materials.

1.2.2 Benefits of multimodal systems

The potential benefits of multimodal presentation systems can be viewed from two different perspectives. These are outlined next.

1) According to the first perspective, benefits derive from the ability to *select a single modality output* or to *switch between modalities* when presentation consists of both visual text and speech outputs. These include:

Freedom of choice: Although the same task may be achieved with equal efficiency using either modality, users may differ in their modality preferences or needs. For example, using speech output with a personalised digital assistant (PDA) device might be unsuited to the environment and not allow privacy for confidential materials. At home, the same application could free the user's eyes so that she could do other things while processing the verbal information. Similarly, in the office, speech output might create background noise for other users, but in a group project, where all people in the room are working on the same task, it could save time. The visual channel might provide backup for further reference and can also enable printing of important information. Multimodality permits the user to switch between modalities according to her preferences or needs.

Multitasking: Different tasks require different forms of presentation as determined by the task context (i.e., by the load placed on each modality by the current interaction). For mobile applications, in which the user's visual attention is occupied, speech output is particularly attractive. Driving a car

is a good example, where visual-spatial information captures the driver's attention. Any additional non-demanding verbal task, such as processing an email message, would be better supported by speech than by visual text. Multimodality enables this flexibility as the user can mentally switch between the two channels.

Perceptual constraints and disabilities: Users may not be able to see, hear, or may not be able to process some types of information easily or at all. A redundant multimodal presentation can accommodate such difficulties either by providing an alternative to the constrained modality or possibly, by allowing one modality to compensate for the deficiencies of the other while using them both. Speech output can compensate for poor viewing conditions (e.g., small screen, shaking train) and poor lighting conditions (under- or over-illuminated rooms). Alternatively, visual output can clarify phonemic ambiguity, as some phonemes can be easily confused (e.g. /m/ and /n/). Similarly, visual output in a noisy environment can facilitate listening (Benoit et al., 1998). For example, when flight connection information is read-out over an in-flight sound system and is backed up by a visual display of text on the airplane's video system. The characteristics of the user might also determine the required output. Text displayed visually benefits users with hearing impairments, whereas speech output is critical for individuals who are blind and for many people with the reading difficulties that often accompany cognitive and learning disabilities (W3C organisation: Web Content Accessibility Guidelines). Moreover, when the user is very busy or tired, listening may be easier than reading (although slower). Finally, young pre-school children using an educational text-to-speech (TTS) application can 'read' a book before they are able to read themselves.

2) Human perception is characterised by the collection and integration of information from different sensory modalities. The second perspective sees the potential benefits of multimodal presentation as deriving from the *concurrent use of both modalities.* The concurrent use of both modalities has been shown to be useful in computer assisted language learning (CALL) titles that enable the user to listen to the pronunciation of visual words or phrases (Ward, 2002), but theoretically, it can take place in all of the above examples. Other than compensating for perceptual limitations, a concurrent use of both modalities may support the cognitive capabilities of the human information processing system so as to reduce user cost. As suggested earlier, the benefit of concurrent multimodality has been shown at the level of individual words (in the form of cross-modal priming studies) and is evident intuitively at the simple message level (e.g., television advertisements and video clips). However, care should be taken due to the potentially negative effects of concurrent multimodality. Cognitive overload, distraction and fatigue have been well documented (c.f., Martin et al., 1988, Wickens, 1992) and they are also evident intuitively, in common everyday situations where, for example, a speaker uses a visual 'Power Point' presentation of visual text to convey complementary or redundant information.

To summarise, multimodal interfaces may support a wider range of diverse applications in various environmental conditions and used by a broader spectrum of the population. The inclusion of both visual and auditory outputs can expand the accessibility of computing for users of different ages, skill levels, cognitive styles and sensory impairments.

Furthermore, the concurrent use of both modalities might also reduce processing if used correctly: incorrect use of multimodality can result in negative cognitive side effects such as cognitive overload, distraction and fatigue. Explicit guidelines are therefore needed to achieve an effective multimodal presentation. The next section examines the state of the art in the research and development of multimodal systems and indicates the lack of guidelines available for the design of audio-visual multimodal presentation of verbal materials.

1.2.3 Current applications, domains of applications and the state of the art in multimodal systems

This section provides several examples of multimodal systems. These serve multiple domains, utilise either natural (pre-recorded) or synthetic speech and feature different levels of design complexity. The multimodal systems reviewed include crafted interactive learning systems, commercial applications made for hand-held devices, as well as sophisticated intelligent systems developed by HCI research laboratories.

One interactive learning system is the "Learning Director" tutorial designed to introduce the fundamental concepts of Director 6, a multimedia authoring tool created by Macromedia in 1997. The tutorial provides natural speech output to supplement static visual text and animated graphics in a crafted manner. Speech output is sometimes added redundantly to the static visual text. In other times, speech output elaborates the visual content and occasionally, it is delivered without supplementary visual text. Speech also provides 'referring expressions' for graphic user interface (GUI) objects (e.g., windows, icons, menus and pointers) and animated demonstrations. It should be noted that speech output has been excluded from any help document of later versions of Director (e.g., Director 7, Director 8).

Development of another multimedia-multimodal system was recently attempted by BT Cellnet in the field of personal communications where mobile phones and palm-sized computers are merging. In this system, speech output was to be fully redundant with the visual text, regardless of the complexity of the presented content: email messages would be read aloud word by word, by means of TTS technology, and displayed dynamically (character by character) on the palm-sized screen. The addition of speech to a dynamic visual output in mobile devices raises a new challenge for interface designers, however no guidelines are available to help in the coordination of dynamic visual text with speech output. No system has reached the market to date.

The Intelligent Project Planner (IPP) is a database queries application on the WWW, developed by the MITRE Corporation. This multimedia-multimodal system can be input via text, mouse, direct manipulation and speech recognition, while its output includes text, graphs, map-displays and speech synthesis (Bayer, Kozierok & Kurtz, 1995). Using a TTS synthesizer, synthetic speech is heard with the visual sentence visible to the user, thereby forming a redundant multimodal presentation of sentence materials. No guidelines were used to assist in the design of this audio-visual presentation.

The Multimodal Multimedia Service Kiosk (MASK) was a prototype system that emerged from a European Commission project (ESPRIT project no.9075, completed in 1996), but never reached production. The system was intended for use in French railway stations to allow users to make travel inquiries, purchase tickets and make seat reservations. Multimodal input included speech and touch and output included speech, text, video and graphics. Speech output in the multimodal presentation was either redundant to the static visual text or provided a further elaboration of the visual output, depending on the complexity of the verbal information. Full redundancy was chosen for very short and simple information. For example, when specifying the required destination, vocal feedback was provided by concatenated speech and the selected destination appeared simultaneously on the screen. For longer messages, cross-modal priming was used: prompts were spoken to the user, accompanied by a visual abbreviation of the spoken message. This visual abbreviation was constantly present on the screen (Lamel, Bennacef, Gauvain, Dartiguesy & Temem, 1998). Lengthy instructions and detailed explanations were provided by means of speech or video output (Dowell, Shmueli & Salter, 1995). The design decisions relating to multimodal presentation in the MASK system were guided by preliminary recommendations, based on varying sentence length (Shmueli, 1994). This work will be described in the next section.

The CUBRICON (Neal & Shapiro, 1991) was a system for Air Force Command and Control. This system enabled the user to interact using spoken or typed natural language and gesture, displaying results using combinations of speech, text, maps, and graphics. For example, the user could ask, "Where is the Dresden airbase?" and the computer would reply in speech, "The map on the colour graphics screen is being expanded to include the Dresden airbase." The computer would then say, "The Dresden airbase is located here," as the Dresden airbase icon and a pointing text box blinked. Interaction management was effected via models of the user and the ongoing discourse that not only influenced the generated responses but also managed window layout, based on user focus of attention.

MAGIC (Multimedia Abstract Generation for Intensive Care) is a multimedia-multimodal distributed system that describes the postoperative status of a patient undergoing Coronary Artery Bypass Graft (CABG) surgery (Dalal, Feiner, Kathleen, McKeown, Pan, Zhou, Hollerer, Shaw, Feng & Fromer, 1996). The system uses knowledge-based techniques for planning and generating briefings in written text, speech, and graphics. The generation of content, the form of media and the determination of temporal relations between media objects are made automatically at run-time. In the example provided by Dalal et al. (1996), the system uses speech and graphics to communicate information about a patient's demographics. Visual text is displayed statically in a tabular form and is highlighted in synchronisation with the spoken words.

In MAGIC, the decision of which media to use under which circumstances is determined by the system's *media allocator*. This component specifies one or more media to express each communicative goal using a simple algorithm based on semantic properties. For example, when handling the patient's demographics information, the media allocator determines that the patient's medical record number (MRN) should be presented by means of static visual text only, while other

details such as patient's age, gender, medical history, the required operation and the surgeon's name may be presented to both modalities in a redundant form. These decisions are based on the task characteristics provided by medical experts.

EMBASSI (Herfet, Kirste & Schnaider, 2001) is a multimodal dialogue assistant which aims at connecting and simplifying everyday technologies. EMBASSI focuses on a car scenario, a home entertainment scenario and an ATM scenario. This state-of-the-art multimedia-multimodal system includes graphic and acoustic output that can be presented on a TV or a PDA GUI. Using sets of renderers, the system's presentation planner takes into account the immediate situative context, the current content, the user preferences and special needs (Elting and Michelitsch, 2001). Verbal materials can be presented by both static and dynamic renderers: static renderers for use in TV GUIs (e.g., static freetext renderer for arbitrary content, static list-text renderer for object lists) and dynamic renderers for use in lower resolution PDA GUIs (e.g., dynamic list-text renderer to present object list item by item). A synthetic speech renderer can be used in either device. Multimodal presentation in the EMBASSI is not necessarily redundant. For example, in the home entertainment scenario, the speech renderer can generate cross-modal references to the list-text renderer ("You can see the title in the list on the TV screen") or to the freetext renderer ("The text on the left gives you a summary of the movie").

This system demonstrates a high level of adaptivity: if the system recognises that the user is not facing the TV screen, the planner will emphasise speech output. On the other hand, long and complex lists of data, such as TV program information, are displayed visually. The rendering capabilities are also sensitive to the physical resource limitations of the output device (Elting, 2002). For example, a comprehensive list of movie titles can be presented by means of static visual text on the 800x600 TV screen. When the user leaves the room with a lower resolution PDA device, the whole list cannot be displayed statically. The presentation planner then can switch to another renderer on the PDA (e.g., dynamic scrolling text and/or speech output). The planner manages content selection and organisation, the selection of appropriate output modalities, the coordinated distribution of information between the output modalities, the realisation of content and the coordination of different output channels into a coherent and cohesive output. Despite this sophisticated framework, no guidelines are proposed to guide the coordination of verbal materials. According to Elting (2002), speech output should be used judiciously with the visual text output to present verbal content; questions of when and how to use a specific multimodal presentation should be inferred from psychological presentation knowledge, which is not embedded within the EMBASSI system.

In summary, this section has reviewed examples of multimodal systems that combine speech output with visual text. Each system represents an attempt to increase the bandwidth of the interaction between users and computers. It is notable that all except one of the systems exists or existed only as a research prototype, the one example of a commercial product (Director 6) was discontinued. The problem might originate in the current perspective of multimodal system development, which is technology-oriented rather than user-centred (Dowell, Life & Salter, 1994). A systematic user-centred approach is needed to inform the design process of sentence presentation in multimodal user

interfaces. This may lead to development of an interface which is acceptable to users, and therefore one which is commercially viable. As evident in this short review, multimodal systems vary in terms of the presented content and its linguistic complexity (from single words to complex verbal information) and also in terms of the realisation and the coordination of their output modalities. This variety raises questions of cognitive compatibility with processing limitations and capabilities of multimodal users. Despite these concerns, this review suggests that only preliminary design recommendations are available to support the user's processing of language in multimodal systems. With the exception of the MASK system, the coordination of written text with speech outputs in all other systems that were reviewed in this section was not informed by explicit design guidelines. Usability research is needed to inform a user-centred approach of multimodal interface design, however, attempts to assess user cost in multimodal systems are few, as now described.

1.3 Usability research: combining speech and text in multimodal systems

This section attempts to review usability research of concurrent multimodal output. As pointed out by Vernier & Nigay (2000), multimodality has mainly been studied as a form of input utilising multiple input devices. The cognitive compatibility of multimodal sentence presentation with the processing capabilities of users has received very little attention and was hardly examined in isolation; most of the studies reported next examined the redundancy of the verbal output in combination with static or dynamic images.

In the first, Nugent (1987) investigated various combinations of printed text, graphics and voice output to provide on-line instructions for the use of an oscilloscope. The combinations involving both pictures and voice (with or without print) yielded better performance than combinations lacking one or the other of those channels. Nugent stated that his results are consistent with Paivio's (1971) dual-coding theory, i.e., a person could alternate between the audio and image codes to more effectively obtain required information. The added print did not impair this advantageous processing.

Different results were obtained by Kalyuga, Chandler & Sweller (1999). These researchers explored the presentation of identical printed and spoken texts combined with diagrams, as students learned about solders composed of different ratios of tin to lead. Three different text conditions were presented with the diagrams: printed text only, spoken text only, or printed and spoken text. Kalyuga et al. reported that students in the spoken text only condition outperformed those in both the printed text only and the printed and spoken text conditions. Thus, the addition of visual text to the speech output impaired processing of the static images.

The combined use of speech and text output for presenting sentences was also investigated in Huls & Bos (1995), In this study, supplemental linguistic output (SLO) generated natural language descriptions of the objects the user was manipulating and the actions she was performing. These descriptions were presented automatically in three conditions: printed text only, spoken text only, or printed and spoken text. In a fourth control condition, the 'no linguistic output' condition, the verbal descriptions were not delivered automatically but had to be requested from the menu by calling the

DESCRIBE function from the menu. The tasks comprised copying, moving, and deleting files. The files were referred to either by name or by definite description, e.g., "Copy the research report by Wim about interaction". The referential definite descriptions were chosen so that they resembled the file names; e.g., the research report by Wim about interaction was named INTER-2. However, to induce errors, in ten cases there were two or more file names in the system that were very much alike. For instance, there was also a research report written by Edwin about interaction which was called INTER-1. In order to correctly execute these tasks, the subjects had to make use of the information conveyed to them through the SLO, or obtain that information using the DESCRIBE function. Results suggest a trade off between the speed and accuracy of user performance. The time taken to perform the task was shortest when output was textual only, followed at a distance by the redundant textual and spoken condition, the no linguistic output condition, and finally the spoken only condition. The smallest number of errors was found in the spoken only condition followed by the textual only condition, the textual and spoken condition, and finally the no linguistic output condition. Thus, the redundant verbal output yielded a fairly fast performance but a large number of errors.

Studies of multimodal presentations of text and animations yield a complex pattern of results. In a study about the formation of lightning, Moreno & Mayer (2000) presented four groups of students with animations and concurrent explanations (e.g., "Cool moist air moves over a warmer surface and becomes heated"). Two groups were presented with an animation and concurrent explanation via speech output or via speech and on-screen text. The other two groups received an animation preceding the explanation via speech output or via speech and on-screen text. It was found that students presented with redundant verbal materials outperformed students who learned with non-redundant spoken verbal materials when the presentations were sequential. For simultaneous presentations of animations and explanations, the opposite was true: a split-attention effect between the on-screen text and the animation occurred and the redundant message impaired rather than helped students' learning.

Craig, Gholson & Driscoll (2002) explored the integration of animated agents into multimedia environments using the same lightning scenario. Three groups of subjects were presented with an animated agent (a computerized character designed to facilitate learning) and concurrent explanations in the following forms: printed text only, spoken text only, or printed and spoken text. Students in the agent spoken-only condition significantly outperformed those in the other two conditions, with no differences between printed text and printed text with spoken narration.

The inconsistent performance reported in these studies may derive from the varying linguistic complexity of the sentences used[1] and from the different configurations of the multimedia-multimodal output. None of these studies controlled for the linguistic complexity of the verbal explanations. In addition, the redundancy of the verbal output was examined in combination with visual images, requiring the subjects to split attention between the visual text and the pictorial output.

[1] The sentences themselves are not provided in these papers.

Under these circumstances, optimal performance was always obtained when the verbal information was presented auditorily. The addition of visual text either impaired performance or rendered it unaffected; however, it never improved performance. It was only in the sequential condition of Moreno & Mayer's (2000) study, when the verbal information was not presented simultaneously with the pictorial information, that the added text improved performance. Therefore, in order to examine the effect of redundancy on the processing of the verbal information itself, there is a need to isolate it from any other output.

The study of Shmueli (1994) meets this requirement. In this study, verbal output conditions were separated from those with pictorial output. The focus of research was on the linguistic complexity of the verbal output and on the configurations of the multimodal presentation itself. Subjects were given a travel inquiry task consisting of nine instructions (sub-tasks) relating to a pre-administered scenario description. The instructions varied in length (but also in the complexity of the mental model required for solving the task) and directed the subjects in the selections they could make (type of travel inquiry, destination, departure time, etc). Contrary to the studies reported earlier, this study manipulated the redundancy of the multimodal output. Two multimodal configurations were used: in the first, the visual text provided a full version of the spoken information, whereas in the second, the cross-modal priming version, only the central noun phrase of the spoken sentence was presented visually. Three of the nine sub-tasks were classified as purely verbal: verbal-easy (9 words sentence), verbal-difficult (13 words sentence) and verbal-control (two short sentences - 7 words and 6 words each, presented without added speech). It was found that for the shorter verbal-easy sentence, presentation type did not affect performance. On the other hand, for the longer verbal-difficult sentence, the cross-modal priming version yielded better performance than the fully redundant version[2]. Due to the limited number of test sentences, and the fact that they varied in their problem-solving complexity in addition to length, these results provide only preliminary data regarding the effect of linguistic complexity on processing redundant speech and text.

A systematic approach is needed to assess the cost incurred by the presentation of language. Specifically, since multimodal systems can theoretically present any verbal content in various forms, such an approach requires:

- A characterisation of content – a model of content which can distinguish variations in linguistic complexity using objective means.
- A characterisation of presentation - taking into account the memory demands of each presentation format.
- A cognitive model of multimodal language processing - assigning processing cost as a function of content and presentation. The model should also consider the characteristics of the user and demonstrate how these may affect processing cost (i.e., should be capable of distinguishing individual differences).

[2] Note that these contrasts were not analysed statistically.

These entities may guide a coherent and complete set of usability studies, providing prescriptive guidelines for multimodal presentation of verbal information. The next section examines guidelines that are relevant for multimodal interface design and outlines the adequacy of current HCI theories in addressing these requirements.

1.4 The adequacy of HCI theories in resolving the research question: a review of the state of the art in cognitive architectures and user models

This section reviews current work on multimodality by the HCI community. It first describes the limitations of current guidelines for combining speech and text output in multimodal interface design and then investigates the available taxonomies and models attempting to provide such guidelines. Specifically, taxonomies and models are examined in terms of their characterisation of content and presentation, enabling to assess their potential contribution in guiding multimodal interface design. Finally, one cognitive theory and two architectures are examined in terms of their explanatory power of multimodal processing of sentence materials for varying contents in varying forms of presentation.

1.4.1 Guidelines

Current design methods do not address properly the increased complexity of multimodal interfaces. HCI textbooks, as well as commercial tools supporting the development of multimodal presentation, provide no advice on how to solve the issue of presentation forms of sentences that vary in their linguistic complexity with respect to cognitive constraints. One way to alleviate the problem is to develop a set of standards, under which speech and text output can be brought together to support the processing capabilities of the user.

A good example is provided by Smith & Mosier's (1986) guidelines for speech output. These prescriptive guidelines comply with users' cognitive constraints while taking into account the linguistic complexity of the presented content. For example:

4.0/28 When using computer-generated speech to provide messages, ensure that those messages are short and simple.

Comments:

1. If a user does not understand a written message, s/he can reread it. That is not the case with spoken messages. Though a REPEAT function might be provided, a better solution is to restrict use of speech outputs for short and simple messages.
2. If a user, who may not be watching a display, must be given long or complex messages, it is probably better to provide a simple auditory signal such as a chime, and then display the messages visually for the user to read. In general, users will understand complex messages better when they see them displayed than when they hear them.

A typical example of guidelines for multimodal output is provided by the W3C organisation. The Multimodal Requirements for Voice Markup Languages document (W3C, work in progress) discusses a markup language that allows an author to write an application that uses spoken dialog interaction in conjunction with a visual interface. Output modalities are each considered in terms of the users' cognitive constraints:

1.3 A characteristic of speech output is its serial nature, which can make it a long-winded way of presenting information that could be quickly browsed on a display. The markup will allow an author to use the different characteristics of the modalities in the most appropriate way for the application.

The scope of the multimodal output addressed by the document is wider than that covered in this work, as the markup language will allow speech output to have different content than that of simultaneous output in another media:

1.31 In a speech plus GUI system, the author will be able to choose different text for simultaneous verbal and visual outputs. For example, a list of options may be presented on the screen and simultaneous speech output does not necessarily repeat them (which is long-winded) but can summarise them or present an instruction or warning.

There is also a specific reference for cases in which both output modalities present the same text:

3.4 The markup language specifies that content is to be simultaneously rendered in speech and other media and that output rendering is further coordinated (i.e., synchronised)... For example, in a news application accessed via a PDA, a browser highlights each paragraph of text (e.g., headline) as it renders the corresponding speech.

The emphasis is on the synchronisation between modalities. No reference is given to the linguistic complexity of the presented content:

4.7.2 The markup language should support finer grained synchronisation. Where appropriate, synchronisation of speech with other output media should be supported with SMIL (Synchronised Multimedia Integration Language) or a related standard. For example, to allow a display to synchronise with events in the auditory stream.

A more comprehensive account is provided by Faraday & Sutcliffe's (1997) cognitive walkthrough method for multimedia evaluation. The walkthrough is founded upon an analysis of cognitive processes and representations, including those that participate in multimodal language comprehension. Based on the analysis, a series of guidelines is provided for the evaluation of the use of media and of presentation techniques. For example, the presentation of speech is assessed in terms of the following set of heuristics:

Guideline SE1 What will be 'attended to' (when a unit would have attention shifted to it by default)? Speech is normally dominant over text, requiring no attention to process and will in general be more alerting. It will be in focus by default.

Caveat: Speech must be processed as it arrives, it will be regularly overwritten within the phonological loop. Beware that irrelevant speech will interfere.

Guideline SE2 What will cause an 'emphasise' effect (when a unit will be attended in preference to other units)? To produce an emphasise effect within speech, its loudness or voice may be varied, or a text presentation unit may also be presented with the same content.

Caveat: Use a text unit presented concurrently with the speech if the content of the speech is important, otherwise, there is a danger of the information being lost due to speech overwrite.

Guideline SE3 What will cause a 'link' effect (when two units are combined by a common referent)? Between speech and text unit, a link' effect requires a common referent between the two propositions, and that they are available concurrently, with the text in focus.

Guideline SE4 What will cause a 'foreground' effect (when one unit gives additional information concerning another)? Between speech and text unit a text based summary or heading will give a 'foreground' effect. It is important that similar content is provided within both units. It should be noted that speech and text will compete in processing and confusion can easily occur. A 'link' effect is required between the two units.

These guidelines aim at predicting what will be attended to in a multimodal presentation and at flagging any potential design problem. However, the recommendations are unaffected by the linguistic complexity of the presented content, which may result in inaccurate predictions. For example, according to these guidelines, as long as both media units are synchronised and the content provided within both units is similar, multimodal presentation will reduce processing cost. Confusion between modalities will occur only when the content provided within both units is dissimilar, not when it is long or complex.

Other work on multimodality includes multimodal models and taxonomies. These aim to support the design of effective multimodal systems by means of systematic investigation of their features. Their accountability for varying contents and varying forms of presentation will be discussed next.

1.4.2 Multimodal taxonomies and models

One prominent stream of research on multimodality is concerned with the classification of different media conducted by Bernsen in the last decade (e.g., Bernsen, 1994, 2001). Bernsen has focused on generating taxonomies of unimodal output (i.e., media output) and input modalities. Bernsen (2001) identifies four super classes of output unimodalities *(super level)*: linguistic vs. non-linguistic, analogue vs. non-analogue, arbitrary vs. non-arbitrary and static vs. dynamic. He then produces all possible combinations of these features and conducts a pragmatic reduction of the set according to the relevance of the unimodality to interface design. The result of this procedure is a set of 20 unimodality classes *(generic level)*. Bernsen further expands the unimodality hierarchy into the *atomic level* which produces 46 unimodality classes. Finally, Bernsen sketches a possible further extension into the *subatomic level* which might be conducted when necessary, depending on the prevailing interface domain. The classification aims to address a much wider research problem than that dealt with by this work: given any particular set of information which needs to be exchanged

between user and system during task performance in context, identify the input/output modalities which constitute an optimal solution to the representation and exchange of that information.

All possible forms of spoken and written language are represented in this taxonomy. Yet, no rules are provided for their successful integration. The taxonomy is intended to help experts make selections using their own experience and judgement. According to Bernsen (2001), information-mapping rules fail to generate advice on which modalities to use in designing interaction for a particular application. Such a rule-based approach cannot successfully cope with the inherent complexity of selecting between thousands of potential modality combinations that are subject to multiple constraints imposed by the task context. Instead, Bernsen provides lists of claims, each based on the declarative and functional properties of a given modality. As an example, he lists 25 properties that are relevant for speech output. Most of these properties do not shed light on the effect of linguistic complexity on processing redundant speech and text; they only relate to the cognitive processing limitations of speech output at a very general level (e.g., "Speech input/output modalities, being temporal (serial and transient) and non-spatial, should be presented sequentially rather than in parallel"). Other properties are more informative (e.g., "Dynamic output modalities, being temporal (serial and transient), do not offer the cognitive advantages (with respect to attention and memory) of freedom of perceptual inspection").

Vernier & Nigay (2000) take this classification one step further and present an output multimodality model. This model assists in the selection of modalities (based on their characteristics) and in combining modalities for the design of a coordinated output interface. In this framework, the multimodalities are modelled as a set of unimodalities that are characterised by Bernsen's unimodality properties (Bernsen, 2001). These can be combined temporally, spatially, syntactically or semantically to form multimodal output. However, the model is restricted to graphic visual multimodalities and thus, does not cover audio-visual multimodal processing.

Elting (2002) presents another multimodality model. This model defines several relations between different media to guide the coordination of multimodal output by the presentation planner of the EMBASSI system, described in the previous section. These include synchronisation between modalities (the coordination of dynamic media in time) and 'referring expressions' to guide the coordination of content between modalities. However, this model also fails to address the effect of linguistic complexity on processing redundant speech and text. As suggested earlier, Elting proposes that instructions of when and how to use a specific multimodal presentation should be inferred from psychological presentation knowledge, which is not embedded within the EMBASSI system.

In summary, although existing models and taxonomies are capable of representing all forms of multimodal output (e.g., Bernsen, 2001), their specifications of the memory demands of different presentation formats should be made explicit. In addition, none of the models reviewed in this section accounts for variations in linguistic complexity of different contents. Linguistic complexity factors should be characterised and objective means should be provided for their measure. Finally, a model of user cognition is needed to explain multimodal processing of verbal information and to assign

processing cost as a function of content and presentation. Three different approaches of user cognition will be considered next.

1.4.3 Cognitive theories and architectures

The general idea of utilising cognitive theory in design has long been an objective of research in HCI. One such theory is Wickens' Multiple Resource Theory (MRT) (Wickens & Kessel, 1980, Wickens, Sandry & Vidulich, 1983; Wickens, 1992) that is famous in its predictions for dual tasks. The MRT argues that instead of one single supply of undifferentiated resources, people have several different capacities with resource properties. Two concurrent tasks will suffer greater interference as the component tasks increase in difficulty (demand more resources) and compete for overlapping resources. Furthermore, the effects of difficulty and resource-overlap interact. The greater the degree of resource-overlap, the more pronounced the effect of the level of difficulty of one task on the level of performance of another task. Resources can be defined by three dichotomous dimensions:

- Two stage defined resources: perceptual and central processing activities versus selection and production of responses.
- Two modality defined resources: auditory versus visual encoding.
- Two processing-code defined resources: spatial versus verbal.

If processing the visual and the auditory information in multimodal sentence presentation can be defined as two separate tasks then, despite the use of both auditory and visual encoding, the MRT predicts a competition for verbal linguistic resources at the central processing level. Moreover, due to the reliance on overlapping resources, increasing either the linguistic complexity of the presented material or the memory demands of the multimodal configuration will increase the interference in processing the multimodal information. In spite of these inferences, the MRT does not deal explicitly with language processing; it was not designed to deal with the concurrent processing of redundant inputs and cannot easily relate to different multimodal configurations. As a result, this theory might be difficult to apply in a manner that can guide the design of effective multimodal systems. A model of user cognition is needed that can explain the combined processes of reading and listening.

One model is the Executive Process-Interactive Control (EPIC) model (Kieras & Meyer 1997). EPIC is a computational performance simulation model designed to explicitly couple perceptual-motor and basic information processing mechanisms. In contrast to the MRT, EPIC does not make the assumption that the overall capacity for cognitive processing is limited. Performance is constrained by structural limitations on perceptual and motor mechanisms and by a limited verbal working memory (WM) capacity. Within this architecture, information flows forward from peripheral sensors through perceptual processors (with distinct processing-time characteristics for each sensory modality) to a production rule cognitive processor (consisting of a production rule interpreter and a WM). Outputs of the cognitive processor control motor processors that move peripheral organs (e.g., moving eye fixation from one point to the other). The WM contains all the temporary information needed for and manipulated by the production rules, including control items (such as task goals and

sequencing information) along with coded sensory input (with separate partitions for auditory, visual and tactile information) and selected motor outputs.

Similar to the MRT, EPIC can be applied to the multiple-task domain. The theory postulates an executive-control process that enables the coordination of multiple tasks by means of production rules. For instance, executive rules can dynamically switch control between different tasks and may control sensory and motor peripherals (e.g., oculomotor processor) directly in order to allocate resources between two tasks. Therefore, if processing redundant information in multimodal sentence presentation can be defined as two separate tasks, then EPIC would predict competition for resources by the executive control process. In addition, the control assigned to this production rule interpreter over motor processors may be used to explain the production of regressive eye movements when reading a difficult piece of text and may be elaborated to explicate multimodal processing of complex text in various multimodal configurations. However, while EPIC can accurately predict the timing of task performance with simple user interfaces (i.e., searching menu displays), it suffers from an increasing burden of configuration as the complexity of the presentation increases. Furthermore, at this time, EPIC is not explicit about capacity limitation and decay of information within WM. This information is required to explain the multimodal processing of sentences that vary in their linguistic complexity and makes the theory difficult to apply in multimodal interface design.

A third model is Barnard's (1985) Interacting Cognitive Subsystems (ICS). The ICS represents the human information processing mechanism as a highly parallel organisation with a modular structure. Its architecture contains a set of functionally distinct subsystems, each with equivalent capabilities, yet each specialised to deal with a different class of representation. These subsystems exchange representations of information directly, with no role for a 'central processor' or 'limited capacity working memory'. The assumption is of a system of distributed cognitive resources, in which behaviour arises out of the coordinated operation of the constituent parts. In this architecture, the acoustic and the visual sensory subsystems transform sensory information into specific mental codes that represent the structure and content of incoming data. These representations are then handled by subsystems that specialise in the processing of higher-level representations: the morphonolexical subsystem for processing the surface structure of language, the object subsystem for processing visuo-spatial structures (including shape of letters and words), and the propositional and implicational subsystems for more abstract and conceptual representations. The output of these higher-level subsystems are directed to the effector subsystems (articulatory and limb).

The ICS is able in principle to deal with the concurrent processing of redundant or complementary inputs which are part of a common task. Its architectural rules allow integration of information in two ways: in direct processing by the blending of information from different sources, and in buffered processing by transformation processes using an extended representation from the image (buffer) record, hence supporting the integration of information over time. An inspection of this architecture suggests that the concurrent use of text and speech may produce competition for processing resources. Competition in processing may take place in the morphonolexical subsystem (at the phonological level) and consequently in the propositional subsystem (at the semantic level), as both

serve visual and auditory input. The morphonolexical subsystem is the first contact point for the cross modal input: acoustic representations from the acoustic subsystem and word-form information from the object subsystem. Competition may arise as a result of different data-flow rates of the two sources of input. Specifically, the ICS suggests that the transformation processes of the morphonolexical subsystem operate at a rate compatible with the flow of transient, acoustically derived, representations. In contrast, in reading statically presented text, the objects are written words which are less transient entities. The products of object-to-morphonolexical transformation, generated in the course of reading, are on the wrong time-base and hence, cannot be handled directly by morphonolexical processes. The mechanism copes with this by using buffered processing. When processing concurrently displayed static text and speech, the morphonolexical subsystem can disengage from processing the auditory input in order to process visually originated representations from the image record. The cost, however, is the separation of the data streams at this level of processing.

Cross modal input may also facilitate performance: Barnard & May (1993) suggest that cross modal integration between acoustic and visual information can take place within the morphonolexical subsystem when the arriving representations are on the same time base and form a coherent data stream. This implies that given that the visual text is synchronised with the spoken information in a redundant multimodal presentation, coherent representation will be formed. It is only when reading is user-paced or when the content of the two input streams differ, that integration processes may cause an interference effect at the morphonolexical subsystem.

The ICS can therefore provide specific predictions for different multimodal configurations. However, this architecture suffers from a few limitations. Similar to the guidelines proposed by Faraday & Sutcliffe (1997), its predictions are unaffected by the linguistic complexity of the presented content. Theoretically, the model may be elaborated to account for processing sentences that vary in their linguistic complexity, but at this time fails to specify integration mechanisms of multimodal information that are required for such elaboration. For example, the architecture does not reveal whether users synchronise between modalities under all load conditions, aiming to optimise the cross modal processing within the morphonolexical subsystem, or whether they use other strategies when load increases. Whereas this minor problem could be easily solved, a more critical issue is the fact that the ICS posits no role for a 'limited capacity WM'. In principle, there is no reason why a distributed model like the ICS cannot assign capacity limits on its various subsystems, yet this critical requirement is more difficult to achieve.

Therefore, although the ICS can serve as a basis for identifying capacity limitations and although it can be elaborated to account for specific integration mechanisms of multimodal input, a simpler option would be to use a language processing framework with a built-in mechanism for explaining capacity limitations, even if such framework does not have a mechanism to explain the separate processes of reading and listening and their convergence. The peripheral processes of reading and listening are assumed to be easier to characterise and relate to an a-modal language processing model than the elaboration of the ICS to account for capacity limitations. The chosen language processing

framework is Just & Carpenter's (1992) capacity theory. Its account of individual differences in verbal WM capacity has been successfully used to explain a wide range of linguistic tasks and may also mediate the processing cost of verbal materials presented to both ear and eye. The capacity theory will be described in detail in Chapter 3.

In summary, this section has reviewed possible platforms upon which to build a cognitive model of multimodal language processing and all of those bases have been found wanting. As an alternative, the thesis is going to adopt Just & Carpenter's (1992) capacity theory as the primary basis, upon which a cognitive model of multimodal language processing will be built. The complete plan of research is outlined next.

1.5 The alternative: a multimodal user model informed by a multimodal design space; an evolution of a cognitive model through experimental validation

The design of a consistently effective multimodal presentation must be grounded in a model of user cognition that can explain the combined processes of reading and listening in various presentation configurations of sentences that vary in their linguistic complexity. As noted earlier, current explanations of multimodal language processing have been limited to single word presentation of speech and visual text. This work, by contrast, aims to explicate the processing of whole sentences presented by visual text and speech modalities in various forms. To achieve this goal, it presents two theoretical entities: a characterisation of the fundamental dimensions of multimodal user interfaces and a multimodal cognitive model of the user.

The multimodal user model (MMUM) attempts to establish a systematic account of individual and common mechanisms for speech and text processing, bringing into a single conceptual structure the established - but hitherto isolated - theories of reading, of listening and of a-modal language comprehension. The model describes the multimodal processing of language at lexical, phonological, syntactic, semantic and pragmatic levels. It attempts to characterise the cognitive structures and processes underlying human language processing in multimodal presentation, including the supervisory attentional mechanisms that coordinate the processing of language in parallel modalities. Moreover, it includes an account of individual differences in WM and can therefore provide specific hypotheses regarding the variation in the user's cognitive (psycholinguistic) cost in response to different configurations of multimodal presentation. Hence both the facilitatory effects and the disruptive effects of multimodality are expressed with respect to user cost in this model. In order to make these predictions, the MMUM is applied to a second theoretical entity, the multimodal design space (MMDS); a space in which all types of redundant multimodal user interfaces can be found. The dimensions of the design space include (i) aspects of the verbal content (as expressed by linguistic complexity), (ii) aspects of presentation (as expressed by forms of media allocation, media realisation and media coordination) and, (iii) user cost. The MMUM takes as input specifications of the content

and its representational form and returns as output to the MMDS a statement of the expected user cost.

The model is validated and refined by laboratory studies with users. The specific aim of these studies is to explore how psycholinguistic processing cost, varying with linguistic complexity, can be optimised through the coordination of multimodal presentation design. Findings of these studies are then translated to a set of guidelines for effective multimodal presentation of sentence materials, thereby informing the design process of multimodal user interfaces.

1.6 Thesis structure

Chapter 2 presents the MMDS. Here the fundamental dimensions of multimodal user interfaces are characterised and their effect on user cost is examined in detail. Linguistic complexity factors are made explicit and objective means are provided for their measure. Also, speech and text are analysed in terms of their *dynamism* and the *durability* that determine possible forms of coordination in redundant multimodal presentation. Specific hypotheses regarding the variation in the user's (psycholinguistic) *cost* in response to different contents and different multimodal configurations (as specified by the MMDS), are provided in Chapter 4 following the description of the MMUM in Chapter 3. The subsequent chapters report a set of studies in which these hypotheses are tested.

Chapter 5 presents the first experiment that investigates the primary assumption of the MMUM: the assumption of synchronised processing of multimodal output. Specifically, the experiment examines the assumed transition from a facilitatory synchronous processing to an interfering asynchronous processing in a static-durable multimodal presentation of *long simple* and *complex* sentences. The second experiment, presented in Chapter 6, investigates the role of *durability* in the comprehension of *simple* and *complex* sentences presented dynamically to both modalities. Chapter 7 describes the third experiment. This experiment investigates the combined effect of the *dynamism* of durable visual text and the *synchronisation* between modalities on user cost, given variations in syntactic complexity. The experiments also investigate the effect of *individual differences* in verbal WM capacity on user cost, given variations in syntactic complexity, multimodality, visual durability and visual dynamism. Chapter 8 translates the experimental findings to a set of guidelines for effective multimodal presentation of sentences and examines the validity of some of them in an applied setting. The work is then concluded in Chapter 9. This chapter summarises the contributions and limitations of this work and evaluates the research process and its products. Possible directions of future work are outlined last.

Chapter 2

The design space of multimodal presentation

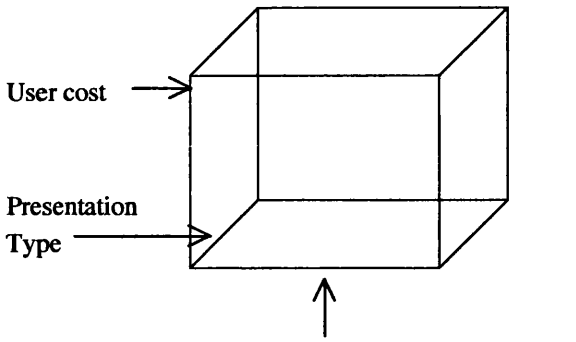
In this chapter, the fundamental dimensions of redundant multimodal interfaces are characterised and their effect on user cost is examined in detail. Following a short description of the multimodal design space (MMDS), the chapter examines factors underlying linguistic complexity and contemplates their selection for further investigation. Objective measures are provided for the quantification of the selected factors. The chapter then looks at different characteristics of media realisation techniques, focusing on their dynamism and durability, as these determine possible modes of coordination between modalities. An investigation of different multimodal configurations follows, looking at different ways of coordination between modalities. Possible measures of user cost in the multimodal domain are outlined last.

2.1 The dimensions of the multimodal design space (MMDS)

Multimodal presentation issues have been characterised by Maybury (1993) as generally concerning content selection (i.e., choosing what to say), media allocation (i.e., choosing which media to say it in), media realisation (i.e., choosing how to say items in a particular media) and media coordination (i.e., choosing how to coordinate multiple media). This design space is concerned with these presentation aspects, with one difference: it does not focus on content selection in terms of its suitability for the task at hand. Rather, sentence materials are treated as predetermined.

As noted earlier, the dimensions of the design space are (i) what is presented (as expressed by linguistic complexity), (ii) how it is presented (presentation type expressed by forms of media allocation, media realisation and media coordination) and, (iii) user cost.

Figure 2.1
A Design Space for Multimodal Presentation: User Cost as determined by Linguistic Complexity and Presentation Type



It is suggested that user's processing cost depends upon the relationships between linguistic complexity (of sentences derived from the content selection process), media allocation, media realisation and media coordination (all represented in the MMDS) and the characteristics of the user (external to the MMDS). To obtain a specific measure of user cost, the MMDS has to be coupled with the MMUM, a theoretical cognitive model of the user. The MMDS describes a variety of verbal contents (via linguistic complexity) and various redundant multimodal presentation types. The MMUM characterises the cognitive structures and processes underlying human language processing in multimodal presentation while taking into account the verbal WM capacity of the user. The model describes how a given content, represented in a given form, is processed by different users and makes specific predictions about the expected level of user cost. In other words, the MMUM takes as input specifications of the content and its representational form and returns as output to the MMDS a statement of the expected user cost. Specific hypotheses regarding the variation in the user's processing cost in response to different contents and different multimodal configurations will be provided in Chapter 4, following the description of the MMUM in Chapter 3.

The next sections examine linguistic complexity aspects and factors underlying the design of a multimodal output.

2.2 Aspects of the content: linguistic complexity

This work proposes a treatment of linguistic complexity that is relevant to any domain concerned with presentation of verbal materials. The study of linguistics seeks general accounts of language, which are independent of particular domains. Conforming to this assumption, it is simply suggested that user cost can be minimised if a given linguistic complexity value of a sentence can be linked to a preferable type of presentation. This proposal necessitates a means of assessing the difficulty of different pieces of text. Aiming to provide an objective means of calculating the linguistic complexity of any given sentence, this section reviews different aspects of linguistic complexity; the lexical,

syntactic, semantic and pragmatic complexity properties that make linguistic materials easier or harder to understand.

The question of what features of a text make the text difficult or easy for the user has been the subject of intensive investigation. One method of assessing the difficulty of text is to apply one of the many measures of reading difficulty. Readability formulas are based on statistical correlations between objectively observable features of texts and the reading level of the readers, as measured by standardised tests. The text properties are usually average sentence length, normally based on samples of 100 words, and an estimate of word difficulty, typically based on syllable length or occurrence on a list of high-frequency words. Such an approach proves too superficial as many factors pertaining to both the text and the readers are not taken into account (see Cutler, 1982 and Rayner & Duffy, 1986).

This research adopts an alternative categorisation of linguistic complexity. Following Gibson (1991), each linguistic structure is associated with a processing cost. The human parser heuristically determines its options based upon evaluation of possible representations with respect to lexical, syntactic, semantic and pragmatic properties, each of which is associated with a weight or cost measured in terms of an abstract unit: the *processing load unit* (PLU). The total cost associated with a linguistic structure is assumed to be determined by summing the weights associated with all the properties of that structure. These properties are defined so that the higher the weight of a structure, the less preferred that structure is.

Processing overload effects are explained by the assumption of the existence of maximum load corresponding to the limited capacity of short-term memory: a structure becomes unacceptable at a particular parse state if the combination of the processing weights associated with its properties at that state is greater than the available capacity. Specifically, Gibson hypothesises that there exists a maximal processing load (in PLUs) that the human parser can maintain. He hypothesises that there exists a constant, K, such that the processor maintains only those structures whose processing load is less than or equal to K PLUs. A structure becomes unacceptable at a particular parse state if the combination of the processing cost associated with its lexical, syntactic, semantic and pragmatic properties at that state corresponds to a load greater than the processing overload constant K, in PLUs.

$$\sum_{i=1}^{n} A_i X_i > K \qquad \text{(unacceptable sentence)}$$

$$\sum_{i=1}^{n} A_i X_i \leq K \qquad \text{(acceptable sentence)}$$

K is the maximal allowable processing load (in PLUs),
X_i is the number of PLUs associated with property i,
n is the number of properties that are associated with processing load,
A_i is the number of times property i appears in that structure.

Following this general characterisation of linguistic complexity, the next sections consider the relevance of the lexical, syntactic, semantic and pragmatic complexity properties to the assessment of linguistic complexity in the multimodal domain.

2.2.1 Lexical complexity

Word frequency

There are a number of databases that enable assessment of the frequency of a given word in English. Two famous examples are the Kucera-Francis written frequency count (Kucera-Francis, 1967) and the frequency count from the London-Lund Corpus of English Conversation (Brown, 1984). Previous research indicates that more attentional resources are required to process low-frequency than high-frequency words. The resource demands of visual word recognition have been most directly assessed within the context of simple recognition tasks while performing a concurrent probe-monitoring task. Herdman & Dobbs (1989) required subjects to perform lexical decisions while monitoring for the presentation of an auditory probe tone. The demands of word recognition were indexed in response to the probes so that longer probe latencies reflected greater resource demands. These results showed that probe latencies were longer during concurrent recognition of low-frequency as compared to high-frequency words. A similar pattern of results was found by Herdman (1992) who used the same probe paradigm with concurrent naming, instead of the lexical decision task. These results show that visual word recognition requires attentional resources and that the demands are frequency sensitive, in that more resources are required to process low-frequency than high-frequency words.

Resource demands of visual word recognition have also been assessed within the context of natural reading. Inhoff, Topolski, Vitu & O'Regan (1993) monitored subjects' eye fixations during first versus repeated readings of a passage. The results showed that fixation duration was longer for low-frequency than high-frequency words during the first reading of the passage. With repeated readings this effect was eliminated. In contrast, repeated readings did not modulate the influence of other factors (e.g., saccade size) associated with perceptual analysis of the text. Repeated reading extinguishes the resource demands of frequency-sensitive lexical processes but does not influence the effects of perceptual analyses.

Resource demands of spoken word recognition have also been assessed in a number of tasks. For example, subjects make lexical decisions more quickly for high-frequency words (Bradley & Foster, 1987; Eimas, Hornstein & Payton, 1990). Syntactic-category decisions are also made more quickly for high-frequency words (Eimas et al., 1990). Finally, when listeners are presented with words containing an ambiguous initial phoneme and are asked to label that phoneme, they tend to choose the one that results in a stimulus forming a high-frequency rather than a low-frequency word (Connine, Titone & Wang, 1993).

Low-frequency words might slow both syntactic processing and semantic interpretation because of the greater time needed to access word class and semantic feature information. The inclusion of such

words imposes a higher processing load than that associated with high-frequency words, irrespective of the format of output. However, it is suggested that unless the text is condensed with low-frequency words, substituting low-frequency words for high-frequency words should hardly affect the comprehension of a text. The speed of accessing word meaning depends highly on the background knowledge of the reader. As suggested by Anderson and Davidson (1988), a text with a lot of unfamiliar words is usually about an unfamiliar topic and it is mainly the lack of knowledge of this unfamiliar topic that makes comprehension difficult. This work uses sentences from mundane domains, rather than expert domains, requiring no background knowledge of the user. Thus, although quantification of word frequency is feasible, the frequency factor is not directly investigated in this research.

Regularity of spelling-to sound characteristics

Seidenberg, Waters, Barnes & Tanenhaus (1984) found that naming a low-frequency word with irregular spelling-to-sound characteristics (e.g., *pint*) is slower than naming a low-frequency regular spelling-to-sound word (e.g., *mint*). This effect is specific to visual-word processing and does not take place during processing of spoken output. Since the main concern is with redundant presentation of verbal materials to both modalities, listening to the word will eliminate the effect. Hence, it is assumed that the regularity of spelling-to sound of a given word is of minor importance. This factor will not be investigated in this work.

2.2.2 Syntactic and semantic complexity

The syntactic level of representation exists as the means of relating a linear string of words to a semantic reading of a sentence. Specifically, sentences represent the relations between predicates and their arguments. For example, the verb 'eat' is a predicate that takes two arguments; one argument is the thing doing the eating (the 'eater') and the other argument is the thing being eaten ('eatee'). The argument 'eater' is called Agent and is the one who performs the action or event described by the predicate. The argument 'eatee' is termed Theme and is the entity which undergoes the action or event described by the predicate. In the sentence *John eats the apple,* the verb 'eat' is the Predicate, *John* is the Agent, and *the apple* is the Theme. These Agent and Theme relations of the noun phrases (NPs) to the verb are called *thematic roles.*

Over the past four decades, many theories of syntactic processing have been proposed to explain aspects of syntactic complexity. Most of these theories have focused on open syntactic requirements as the major source of syntactic complexity (e.g., stacking incompletely parsed phrase-structure rules, holding incomplete syntactic/thematic dependencies). Gibson's (1991) complexity metric is based on incomplete thematic dependencies and provides a sophisticated and successful account. In this framework, structure building consists of looking up the current word in the lexicon and then matching the categories of these lexical entries to the predictions in the structures built thus far. Structural integrations considered by the syntactic parser are limited by the syntactic constraints

associated with each lexical item[3]. Specifically, Gibson (1991) hypothesises that there is a memory cost associated with each incomplete dependency involving a thematic-role assignment. He defines a set of local thematic violation constraints, each of which rules out linguistic representations lacking a necessary property. These constraints are designed to reflect the Theta (θ)-Criterion and Projection Principle from Government-Binding (GB) Theory (Chomsky, 1981), the syntactic theory of processing assumed by Gibson (1991):

- The Projection Principle: Lexical requirements must be satisfied at all levels of representation (paraphrased from Chomsky (1981) p. 29).
- The θ-Criterion: Each argument bears one and only one θ-role and each θ-role is assigned to one and only one argument (Chomsky (1981) p. 36).

The second part of the θ-Criterion – that each θ-role be assigned – follows from the Projection Principle so that the rule can be simplified:

- The θ-Criterion (simplified): Each argument bears one and only one θ-role.

Gibson's property of Thematic Reception (TR) associates processing load with this principle:

- The Property of Thematic Reception: Associate a load of X_{TR} PLUs to each Confirmed (C)-node constituent[4] that is in a position that can receive a thematic role in some coexisting structure, but whose θ-assigner is not unambiguously identifiable in the structure in question.

Associating processing load with the Projection Principle gives a similar property. This property is stated in terms of thematic elements. Following early work in linguistic theory, Gibson (1991) distinguishes two kinds of categories: *functional* categories and *thematic* or *content* categories. Thematic categories include nouns, verbs, adjectives and prepositions. Functional categories include determiners, complementisers and inflection markers. He hypothesises that thematic elements are more visible to the parser than their functional counterparts. This assumption is made explicit in the property of Lexical Requirement (LR):

- The Property of Lexical Requirement: Associate a load of X_{LR} PLUs to each lexical requirement that is obligatory in some coexisting structure, but is satisfied by an Hypothesised (H)-node constituent[5] containing no thematic elements in the structure in question (i.e., structures that mark functional categories).

In addition, Gibson assumes that thematic roles seek out categories that contain thematic content. He assumes that a thematic role which is assigned to a semantically null category (e.g., the complimentiser *that*) is passed on to the argument of that category:

- The Property of Thematic Transmission: Associate a load of X_{TT} PLUs to each semantically null C-node category in a position that can receive a thematic role, but whose lexical requirements is currently satisfied by a hypothesised constituent containing no thematic elements.

---

[3] For the description of the syntactic parser postulated by Gibson (1991), see Chapter 3, section 3.2.3.

[4] A confirmed structure built for a given word (see Chpater 3).

[5] Structures predicted to the right of a given word (see Chapter 3).

Each of these properties is associated with a numeric memory cost. Specifically, Gibson (1991) assumes that the loads associated with the property of thematic reception (X_{TR} PLUs) and the property of lexical requirement (X_{LR} PLUs) are the same (since the thematic criterion and projection principle they derive from, are believed to follow from a more general principle, that of local uninterpretability):

X_{TR} PLUs = X_{LR} PLUs = $X_{Interpretation}$ (X_{INT}) PLUs

This assumption does not extend to the property of thematic transmission (PTT)[6]:

X_{TT} PLUs < X_{INT} PLUs

According to Gibson's (1991) theory, the maximal memory capacity is four local thematic violations ($K \leq 4X_{INT}$ PLUs), so that sentences with parse states requiring five local thematic violations ($5X_{INT}$ PLUs) are unprocessable. Consider the following sentences:

Right-branching: The dog bit the woman that likes the man that eats red meat.

Doubly-embedded: The man that the woman that the dog bit likes eats red meat.

The right-branching sentence imposes an acceptable processing load on the user. This structure allows an immediate assignment of thematic roles so that processing load is maintained minimal throughout sentence computation: the processing load associated with the parse of this sentence starts at X_{INT} PLUs when the NP *the dog* is input. When the verb *bit* is input, a thematic role is assigned to the NP *the dog*, but the total load remains X_{INT} PLUs since the verb *bit* has lexical requirements that are yet unsatisfied. When the NP *the woman* is input, the processing load decreases to $0X_{INT}$ PLUs since all principles of interpretation are satisfied by the current input string. The load then goes up when the complementiser *that* is input, marking the initiation of a relative clause. When the next verb *likes* is processed, the lexical requirements of the complementiser are satisfied but the total load remains X_{INT} PLUs since the verb *likes* requires a noun phrase complement, which is not yet satisfied. The load decreases further when the object NP *the man* is input and the cycle of minor load fluctuations repeats itself with further input. The processing load associated with this structure never gets to be greater than X_{INT} PLUs.

The doubly-embedded sentence imposes an excessive processing load on the user. According to Gibson's (1991) complexity metric, this sentence structure includes three lexical NPs as well as two non-lexical NPs, operators, that need thematic roles but lack them. In addition, the second complementiser (*that*) has an unsatisfied lexical requirement:

[$_{IP}$ [$_{NP}$ The man$_i$] [$_{CP}$ [$_{NP}$ O_i] that [$_{IP}$ [$_{NP}$ the woman$_j$ [$_{CP}$ [$_{NP}$ O_j] that [$_{IP}$]]]]]]

X_{TR} X_{TR} X_{TR} X_{TR} X_{LR}

[6] The mathematical justification for this inequality is given in Gibson (1991).

The total load associated with the doubly-embedded structure is $4X_{TR} + X_{LR}$ PLUs = $5X_{INT}$ PLUs, an unacceptable processing load so that processing breakdown will occur[7].

Finally, Gibson proposes two competing principles to determine the exact processing breakdown points in complex sentences:

- The Node Pruning Hypothesis (NPH): If a structure requires more processing load than the available capacity, then prune the entire structure from further consideration.

For the doubly-embedded sentence structure, the NPH principle predicts that parsing should stop when the second instance of *that* is input:

[*IP* [*NP* The man] [*CP* [*NP Oi*] that [*IP* [*NP* the woman$_j$ [*CP* [*NP Oj*] that [*IP*]]]]]]
X_{TR} X_{TR} X_{TR} X_{TR} X_{LR}

---

[7] In 1998, Gibson suggests that his 1991 theory cannot explain discourse-based effects. For example, it cannot explain why the doubly-embedded sentence structure is easier to process when a first- or second-person pronoun (an *indexical* pronoun) is in the subject position of the most embedded clause, as compared with similar structures in which a proper name, a full NP or a pronoun with no referent is in the subject position of the most embedded clause:

(a) Indexical pronoun: The student who the professor who **I** collaborated with had advised copied the article.

(b) Short name: The student who the professor who **Jen** collaborated with had advised copied the article.

(c) Full NP: The student who the professor who the **scientist** collaborated with had advised copied the article.

(d) No referent pronoun: The student who the professor who **they** collaborated with had advised copied the article.

In an acceptability questionnaire, participants rated the items with the indexical pronouns significantly easier to process than any of the other three conditions. He proposes the Syntactic Prediction Locality Theory (SPLT). According to this newer metric, syntactic complexity should be explained by the quantity of computational resources involved in the processing of sentence structures. Specifically, the availability of computational resources is influenced by two components of sentence comprehension: (1) a memory cost component which dictates what quantity of computational resources are required to store a partial input sentence and, (2) an integration cost component which dictates what quantity of computational resources need to be spent on integrating new discourse referents into the structures built so far. Since the referent of a first-person pronoun "I" is in the current discourse (current discourse is assumed to always include a speaker/writer and a hearer/reader), it is not considered to be a new discourse referent and does not increase processing load. Although this theory has a greater explanatory power than the 1991 metric, it was decided that the old metric is sufficient for the purpose of this work. The doubly embedded structure that was used extensively in this research always used a full NP for the subject position of the most embedded clause, and thus correctly reflects the unacceptability of $5X_{INT}$ PLUs structures.

An alternative to the NPH consists of maintaining part of the overly expensive structure. This alternative is based on a heuristic of partial structure removal proposed by Frazier (1985):

- The Least Recent Node Hypothesis (LRNH): If a structure requires more processing load than the available capacity, then selectively remove (forget) nodes directly dependent on the least recent words in the input string until the load associated with the structure is lowered below the desired threshold.

The LRNH principle predicts that people will alter the doubly-embedded structure by discarding the partial structures directly dependent on the least recent words in the input, until the load associated with the structure is less than the maximum available. Thus, the noun phrase (NP) structure headed by the *man* is discarded. Furthermore, the complementiser phrase (CP) modifying this NP is also discarded, since its existence in this structure is dependent on the existence of the head NP:

[*IP* [*NP* the woman$_j$ [*CP* [*NP Oj*] that [*IP*]]]]]]
XTR XTR XLR

The load associated with this structure is below the maximum allowable load, so that processing can continue. All proceeds without difficulty until the verb *eats* is input.
The verb *eats* cannot attach anywhere in this structure and the parsing fails:
[*IP* [*NP* the woman$_j$ [*CP* [*NP Oj*] that [*IP* [*NP* the dog] bit]]] likes]

The prediction of the LRNH has been shown to have a better psycholinguistic validity. For example, it supports Frazier's (1985) finding that ungrammatical sentences like (a) are often accepted as grammatical, while grammatical sentences like (b) are rejected:
(a) *The patient the nurse the clinic had hired met Jack.
(b) #The patient the nurse the clinic had hired admitted met Jack.

The LRNH principle was thus selected to predict the processing breakdown point in this research (for a full discussion of the two principles, see Gibson, 1991).

2.2.3 Sentence length

According to Gibson (1991), sentence length is not a major factor in determining the linguistic complexity of a sentence. He suggests that PLUs are motivated (rather than correspond directly) to units of linguistic short term memory, in that they give the parser a metric upon which to evaluate linguistic structures. This indirect link between PLUs and short-term memory suggests that sentence length does not increase user processing load[8]. For example, excluding the last clause at the end of

[8] According to Gibson's SPLT theory (Gibson, 1998), both the memory cost component and the integration cost component are influenced by the notion of locality. When storing a partial input sentence, the longer a predicted syntactic category is maintained in memory, the greater is the cost for keeping this category in memory. When integrating new words into the structure, "the greater the distance between an incoming word

the right-branching sentence structure: *The dog bit the woman that likes the man that eats red meat* results in a shorter sentence with the same processing load in PLUs: *The dog bit the woman that likes the man.*

However, it seems reasonable to suppose that even if additional length does not impose additional parsing demands, as Gibson suggests, an increased number of thematic roles always adds to the complexity of the propositional content of a sentence. The propositions, constructed as a sentence is parsed, must be retained and progressively integrated with each other. Some important evidence indicates that propositions are retained quite well and much longer than information about the phonological and syntactic structure of a sentence (Sachs, 1967). Thus, it might be thought that all propositions derived from a sentence would be preserved. However, even in Sachs's study in which a recognition procedure was used, performance was not perfect when rejecting sentences that were semantically different from the original even at short delays between presentation and test.

Sentence length is crucial when the task requires the subject to remember the exact wording of a sentence for a period of time before making a response (Levy, 1978; Waters, Komoda & Abuckle, 1985, Martin, 1990). If a sentence remains visible until a response is made, then any forgotten information can be collected again using regressive eye-movements. In this case, the length of the sentence is of minor importance. However, delivering a long sentence by any transient mode of presentation (e.g., auditory presentation, rapid serial visual presentation) or removing a sentence from sight prior to response should affect long sentences more strongly than short ones (c.f., Miyake, Carpenter & Just, 1994).

2.2.4 Pragmatic complexity

Pragmatic information can assist syntactic parsing of sentences (e.g., Ferreira & Clifton, 1986; King & Just, 1991). For example, the doubly-embedded structure *The man that the woman that the dog bit likes eats red meat*, consists of four nouns: *man, woman, dog* and *meat.* Determining which noun is the agent of which of the three verbs *bit, likes* and *eats* is assisted by expectations of the usual relationships found between these nouns. Our pragmatic knowledge includes the animacy of nouns. For example, the noun *meat* can only be the recipient of the verb *eats,* rather than its agent. In addition, our mental model of the world includes a strong association of *dog* as the agent of *bit.* On the other hand, the association between the verb *bit* and the theme *man* is equal in strength to its association with the theme *woman.* Because pragmatic associations between nouns and verbs are independent of syntactic relationship, this kind of information is potentially misleading.

---

and the head or dependent to which it attaches, the greater the integration cost" (Gibson 1998, p. 8). This relates only indirectly to sentence length, as syntactic categories will be maintained longer in memory only in long sentences; similar to the 1991 metric, the newer theory claims that the length factor itself does not increase processing load.

A pragmatic complexity metric is provided to determine the ease of interpretation based solely on a pragmatic reading of complex sentences. An example, based on the doubly-embedded and the right-branching structures, follows. In this simple metric, letters code for the sentence's four nouns whereas numerals code for the three verbs:

The dog bit the woman that likes the man that eats red meat.
A 1 B 2 C 3 D

The metric assumes that the correct relationship between 3-D (*eats red meat*) is irreversible. All possible combinations of thematic roles in this sentence are presented next. The correct sets are marked with (T). Pragmatically meaningless items are marked with (-). Reversible options are marked with (R). The possible relationships between A, B, C, D and 1, 2, 3 are as follows:

A1B. The dog bit the woman (T)
R. The woman bit the dog (-)
A2B. The dog likes the woman (F)
R. The woman likes the dog (F)

A1C. The dog bit the man (F)
R. The man bit the dog (-)
A2C. The dog likes the man (F)
R. The man likes the dog (F)

B1C. The woman bit the man (F)
R. The man bit the woman (F)
B2C. The woman likes the man (T)
R. The man likes the woman (F)

A3D. The dog eats red meat (F)
B3D. The woman eats red meat (F)
C3D. The man eats red meat (T)

The total count yields fifteen possible sets, of which three are correct and two are pragmatically meaningless. Given an equal probability to assign any verb to any noun[9], the pragmatic plausible sentence-space is thirteen. It is suggested that this sentence is easier to understand than a sentence with a higher pragmatic count and with an identical value of syntactic complexity (e.g., *The banker that the lawyer that the couple met represented went bankrupt*).

2.3 Aspects of multimodal presentation

As suggested earlier, the proposed account of linguistic complexity is relevant to any task that is concerned with presentation of verbal materials. However, different tasks require different forms of

[9] The assumption that the probability to assign any verb to any nouns is equal will be proved over-simplified in the continuation of the thesis. It appears that when full NPs (common nouns) are used, subjects use pragmatic knowledge, with associations varying in strength, to assign thematic roles to the sentence constituents. It is only when proper nouns (names) are used that the assumption proves correct (see Chapter 8 for a further discussion).

presentation as determined by the task context (i.e., by the load placed on each modality by the current interaction). A few examples were described in the previous chapter. For example, a unimodal speech output was claimed to serve best any secondary verbal task while driving. Other examples suggested the combination of the two modalities, as in systems that provide redundant speech output to supplement a static visual presentation of text (e.g., MAGIC, EMBASSI). Finally, task context was also claimed to sometimes require a display of text on a low-resolution device such as mobile phones, pagers and palm pilot (e.g., BT Cellnet, EMBASSI), raising a different challenge for multimodal interface designers. Common to these examples is the requirement for a systematic investigation of media allocation, realisation and integration aiming to maximise the bandwidth of the interaction.

2.3.1 Media allocation and realisation

Natural language can be presented visually (i.e., visual text) and auditorily (i.e., spoken language). The allocation of verbal sentences to these two modalities should include a consideration of their physical and time-dependent characteristics.

Spoken language:

Physical characteristics: Physically, the stimulus of speech is a continuous variation of oscillation of air pressure reaching the eardrum. The human ear is capable of various range of intensities (between 90Db and 150Db) and frequencies (between 20Hz and 20,500Hz). The frequency content of articulated speech changes rapidly and systematically over time. In addition, different voices have different pitches (female voices have a range of 300Hz to 400Hz whereas male voices have a range of 150Hz to 250Hz), (Wickens, 1992).

Previous research has demonstrated that synthetic speech is less well recalled than natural speech. Luce, Feustel & Pisoni (1983) concluded that this is because synthetic speech increases the effort involved in encoding and/or rehearsal of presented information. Waterworth & Thomas (1985) examined an ordered recall of lists of ten words spoken in either a synthetic or a natural voice, and used repetition of the words as a measure of successful encoding. The results indicate that most of the memory deficit with synthetic speech is due to encoding difficulties, rather than problems with item retention. The experiments provide evidence that encoding synthetic speech involves more processing capacity than does encoding natural speech, but that once it is encoded it is stored just as efficiently.

Time-dependent characteristics: Human speech perception is rapid, continuous, and incremental: due to its *transient* and *dynamic* nature (c.f., Bernsen, 1994, 2001), speech must be processed as it arrives. Furthermore, as a dynamic media, speech attracts user attention automatically. It cannot be ignored, either at the phonological level (c.f., the unattended speech effect), or at the semantic level; background meaningful speech impairs performance of a concurrent reading task (Martin, Vogalter & Forlano, 1988). Moreover, due to its dynamic nature, speech presentation rate should be

controlled. For example, if the user has to extract a complex message from speech, then the pace of presentation may over-run her ability to comprehend and memorise the message. A presentation rate of 160 words per minute is considered to be legible.

Spoken natural language does not allow the same flexibility provided by a static visual text presentation. In gaze-duration studies, readers sometimes initiate regressions to earlier portions of the sentence when trying to interpret a later part. For auditory comprehension, any regression would have to be carried out in the mind of the listener, as speech must be processed in real-time. Also, perceptual ambiguity and ambiguity in lexical segmentation may also increase memory demands.

Prosody and intonation (or the rhythm and melody of spoken language): Spoken sentences contain rich cues to syntactic structure in the form of prosodic information. For instance, the direction of pitch-change on a single word can signal whether that word is uttered as a statement (with falling pitch) or as a question (with rising pitch). Furthermore, words carrying important or new information tend to be uttered more loudly than those carrying unimportant or old information. Further functions may be carried by the lengthening of a word in a sentence, which may signal an important structural break and also by pauses, which signal major syntactic boundaries (Cooper, summarised in Cooper & Paccia-Cooper, 1980). Ferreira (1991, 1993) demonstrated that prosodic information is not the perfect predictor of syntactic structure but that the patterns are regular enough to be potentially useful to the comprehension system.

Visual text:

Physical characteristics: Text formats can vary in size (point), colour and shape (font, bold, italics). In addition to these generic properties, there are media-related properties that seem to influence the ease of reading. For example, printed and electronic texts are not identical in terms of their physical characteristics. Muter, Latremouille, Treurniet, & Beam (1982) compared speed and comprehension in reading from a videotex terminal and a book. Results over two hours of reading indicated that, though extended reading from videotex was feasible, it was 28% slower than reading from paper. There was no significant difference in comprehension. Several other researchers have also found decreased efficiency from CRTs of the 1980s (e.g., Gould & Grischkowsky, 1984; Wilkinson & Robinshaw, 1987), some with reading and some with proofreading. There are many typical differences between book and computer reading that conceivably could account for the observed slower reading from computer screens of the 1980s. It is quite clear that no single variable accounts for the obtained differences in performance between CRTs and paper. Several variables, including resolution, interline spacing, polarity, and edge sharpness contribute to the effect (Gould, Alfaro, Barnes, Finn, Grischkowsky, & Minuto, 1987; Kruk & Muter, 1984; Muter & Maurutto, 1991). With a more modern system, including a large higher-resolution screen with dark characters on a light background, reading from a computer screen can be as efficient as reading from a book (Muter & Maurutto, 1991).

Time-dependent characteristics: Visual text can be presented by means of either *static* or *dynamic* displays (c.f., Bernsen, 1994, 2001). A static display does not involve the dimension of time during

the presentation itself since effectively the full text appears simultaneously on the screen. Note however that a time dependent characteristic of static visual text is the duration of presentation, in case that the text is removed after a period of time (in this work, the static text is removed after a period equating to a predefined reading rate period). This presentation form enables faster processing of pieces of text than does speech presentation; average reading rates are about 200–300 words per minute, although they can be lower for difficult pieces of text (c.f., King & Just, 1991). Maximal reading rates (achieved by fewer fixations, shorter fixation durations and fewer regressive eye-movements) are estimated at about 800–900 words per minute (Masson, 1985). In addition, when reading a static visual text, some of the memory demands of auditory comprehension are avoided. For example, in moderately complex sentences (whose processing load is lower than *K*), there would be no need to retain a surface form of a sentence downstream from the point at which syntactic complexity occurred because the remainder of the sentence would remain available on the display. The availability of the text for further processing is expected to increase the number of regressive eye-movements to earlier parts of the text. Regressive eye-movements take about 10-15% of the time when reading typical texts (Rayner & Pollastek, 1987). However, they are affected by the syntactic complexity of the text (Ni, Shankweiler and Crain, 1996), by local syntactic ambiguities in the text (Frazier & Rayner, 1982), by semantic anomalies in the text (Rayner, Carlson & Frazier, 1983) and by reading skill (Posner, Abdullaev, McCandliss & Sereno, 1997).

A restricted (low-resolution) display may require a *dynamic* presentation of visual text. Two methods of dynamic text presentation that have been tested are rapid serial visual presentation (*RSVP*) for isolated words and the *Times Square Format*. With RSVP, text is presented at a fixed location on the screen, one word at a time or a few words at a time and the eye processes successive information using fixed gazes. One consequence of reading with the RSVP technique is that regressions to earlier parts of the text are effectively eliminated. Several researchers have demonstrated that readers can perform approximately as efficiently with RSVP as with normal page-format reading (e.g., Juola, Ward, & McNamara, 1982). In addition, Kang & Muter (1989) found that smooth (pixel-by-pixel) horizontal scrolling with a small window (Times Square Format) produced performance at least as good as RSVP, contrary to earlier studies which did not use pixel-by-pixel scrolling. Furthermore, subjects preferred the Times Square Format, which is often used in electronic billboards. Due to the transient nature of these dynamic techniques, they both impose a higher load on the user than the presentation of a static-durable text when the text is long or complex (Öquist & Goldstein, 2003). Since the visual text must be processed as it arrives, presentation rate should be controlled as in speech presentation.

A dynamic text presentation can also be *durable*. Text can appear one character or one word at a time, starting in the left corner of the screen. Since dynamic media attracts attention automatically, this may impair the flexibility to perform regressive eye-movements enabled by a static text presentation. Also, such presentation methods do not allow the utilisation of parafoveal viewing. Word identification information is obtained foveally and parafoveally. Readers utilise partial word information from parafoveal vision (Inhoff & Rayner, 1986; Pollastek, Rayner & Balota, 1986; Rayner, 1975) as indicated by shorter fixation times when the first three letters are available on prior

fixations (Inhoff, 1985; Lima, 1987). In the absence of such partial word information, reading rate is slowed approximately by one third (Rayner, Well, Pollastek & Bertera, 1982). On the other hand, Tombaugh, Arkin & Dillon (1985) established that the comprehension of a dynamic presentation rate of 30 characters per second equals the level obtained for a static presentation.

2.3.2 Media coordination

The MMDS only addresses fully redundant multimodality - the presentation of completely identical contents in both modalities. Other forms of multimodality are not assessed by this work. Using the terms devised by Vernier and Nigay (2000) for their classification scheme of output multimodality, these include:

- Complementary multimodality - when the meanings conveyed by the two modalities are complementary. For example, a designer of a home entertainment multimodal system can choose different texts for simultaneous auditory and visual outputs: a comprehensive list of movies may be presented on the TV screen and the simultaneous speech output would include an instruction to select the required title.
- Complementary-Redundant multimodality - when the information conveyed by the two modalities is partially redundant and complementary. Using the above example, a list of movie titles and their corresponding presentation times may be presented visually on the screen and a description of each title is provided by speech.
- Partially-Redundant multimodality - when one modality conveys a subpart of the information that the second modality conveys. For example, the use of cross-modal priming in the MASK system: prompts are spoken to the user and are accompanied by a static visual abbreviation of the spoken message.

In spite of the limited focus of the MMDS, a fully redundant multimodal presentation of verbal materials involves different forms of coordination between the visual and the auditory modalities. This section is concerned with these different forms. In the previous section, limited degrees of freedom were suggested for the realisation of the speech media: only physical features, presentation rate and prosodic features can be controlled. On the other hand, visual text may be presented in a *static-durable* form, a *dynamic-transient* form or a *dynamic-durable* form. The coordination of the multimodal presentation depends on the realisation of the visual text in one of these three forms.

When the visual text is presented in a *static-durable* form, the addition of speech results in a non-coupled (*asynchronous*)[10] presentation of the verbal materials. The users can determine the synchrony

[10] The concept of synchronisation has two different meanings. Synchronisation relates both to the form of coordination between modalities and to the mode of processing of the multimodal information. To minimise confusion in its use, when the two modalities are presented together at the same rate, presentation is termed 'coupled'. When the visual text is static, presentation is termed non-coupled. The term 'synchronisation' is used to describe a specific case of coordinated processing between modalities and relates to the user's focus of

in such a non-coupled presentation through coordinating their reading with their listening. This presentation form has the advantages of allowing the user to scan and to skim the visual text. It also enables the user to perform regressive eye-movements to previously processed portions of text. On the other hand, the dynamic and transient properties of speech attract attention automatically and yield an automatic processing of the spoken message.

When the visual text is presented in a *dynamic-transient* form (e.g., RSVP), presentation of speech may be coupled (*synchronous*), given that the outputs of the two modalities occur at the same time[11]. If the message of the auditory modality precedes or comes later than the visual message, then the multimodal presentation is "out of phase"[12]. Even slight discrepancies are noticeable such as when a film has been dubbed and the speech no longer matches the movements of the speaker's lips.

Finally, when the visual text is presented in a *dynamic-durable* form, for example in a word-by-word accumulating text, a redundant presentation of speech may be coupled (*synchronous*) but allow regressive eye-movements to previously read portions of text.

It is suggested that these modes of coordination differ in terms of two central factors (i) the memory demands imposed by the multimodal presentation types and (ii) the use of synchronised processing in these presentation forms. The potential effect of these factors on user's processing cost will be considered for various verbal contents in Chapter 4, following the description of the MMUM in Chapter 3.

2.4 Assessment of user cost in the multimodal domain

Assigning a preferable presentation to a given linguistic complexity value of a sentence will optimise user cost, the third dimension of the MMDS. The question is how to measure user cost in multimodal processing. Ergonomics and HCI have traditionally assessed user's processing cost of verbal information in terms of intelligibility, comprehensibility and time and error metrics. This section summarises the methods and metrics that have been most widely used to assess user cost in the visual and the auditory domains and suggests that some are domain specific whereas others are not applicable for the assessment of the user's (psycholinguistic) cost in the multimodal domain.

Readability formulas have been mentioned earlier in this chapter. These prescribed metrics rely on statistical correlations between objectively observable features of text and reading levels of readers. As suggested earlier, these formulas are too superficial, as many factors pertaining to both the text and the readers are not taken into account. Further, the measurements that enter into formulas are often inaccurate reflections of the difficulty of text. Another measure of text difficulty, similar in

---

attention. Specifically, synchronous processing requires a concurrent focus of attention on the same word in the two modalities.

[11] See footnote 10.

[12] This form of presentation will not be investigated in this research.

some respects to a readability formula, but this time requiring readers, is the Cloze Procedure (Taylor, 1953). Using this technique, samples of passages are presented with every *n*th word missing, and readers are required to fill the missing gaps. The Cloze Procedure has an advantage over readability formulas in that it can be used to assess the effects of the presence of other features (such as illustration or underlining) on the comprehension of text (See Hartley, Bartlett & Brantwaite, 1980; Newton, 1983; Reid, Briggs & Beveridge, 1983). A rather different measure of text difficulty is to ask the readers to circle areas in the text where they find words or sentences difficult (Hartley, 1995). These last two measures can be used at the stage of designing a text and determining its content. However, they cannot contribute to the aims of this research.

In the auditory domain, the assessment of user cost mainly concerns the effect of the physical characteristics of the speech output on perception and comprehension. The standard methods for assessing speech are traditionally divided into methods for objectively assessing speech intelligibility and subjectively assessing speech quality. The most common established methods for assessing speech intelligibility require the subjects to listen to syllables, words or complete sentences. Subjects are asked to write down what is heard or to answer comprehension questions relating to a passage of spoken text. Such measures have commonly been employed for the quality assessment of synthetic speech (Kryter, 1972). Perceived speech quality is ascertained by having listeners rate speech quality. The speech that is assessed in these tests is generally specific material, recorded or spoken under defined conditions. Two common scales are *the 5-point listening-quality scale* and the *listening effort scale* - another 5-point scale, which requires the listener to rate the ease with which the meaning of the speech is understood.

It is claimed that these approaches provide rather practical and domain specific measures. In contrast, the variety of contents (from single words to complex verbal information) and the variety of multimodal presentation types raise the need for a more elaborate approach to assess the cost. To answer questions of cognitive compatibility with processing limitations and capabilities of multimodal users, the required approach needs to take into account both the linguistic complexity of the verbal material in the form of prescribed metrics and the memory demand incurred by the multimodal presentation type. In addition, it should consider the verbal WM capacity of the user, as this might affect user cost. Furthermore, user cost needs to be made explicit in the context of multimodal research using on-line measures of attention and comprehension.

This suggests the need to revert to measures used in cognitive psychology for assessing user cost. The two dependent variables most commonly used have been reading times and comprehension rate, as measured by recall, questionnaires or error detection. Reading times can measure an absolute time of processing a sentence or the time it takes to process specific words or clauses. The eye movement monitoring technique provides measures of location and duration of eye fixations. It enables the determination of a moment-by-moment profile of processing load across different sentences presented by means of durable text. Unfortunately, it does not enable to compare between durable and non-durable presentation forms.

The moving-window technique (Just, Carpenter & Woolley, 1982) is less costly and appears to be reasonably sensitive to processing load across visually presented sentences. This task works as follows: a trial begins with every letter of a sentence concealed but indicated by a position marker. The subjects begin by pressing a pacing button to reveal the sentence's first word. Once the word has been comprehended, the subject presses the pacing button again. The button press simultaneously conceals that word and reveals the next one. The subject proceeds in this manner until the end of the sentence. A subset of the sentences is typically followed by a question to ensure that the subjects read for comprehension. A similar technique was developed by Ferreria, Henderson, Anes, Weeks & McFarlane (1996) to assess the processing load across auditorily presented sentences. The auditory moving window technique allows participants to listen to the sentence one word at a time by pressing a pacing button to receive successive words. Times between button presses are then recorded. Despite their capability to assess user cost throughout sentence computation, these techniques are not suitable for the purposes of this research. The experiments aim to assess two modes of multimodal presentation: a coupled presentation and a non-coupled presentation. In the non-coupled presentation, the visual text is presented in a static-durable form. The visual moving-window technique requires a transient-dynamic presentation of the visual text. If the visual text has a static-durable form, subjects may find it difficult to focus on the text while pressing a pacing button to receive successive auditory words. Such a procedure would be expected to impair any form of divided attention between the two modalities. On the other hand, a coupled multimodal presentation enables use of a single pacing button to receive a synchronised multimodal output. However, such a procedure would be expected to impair the intelligibility of the spoken output and is thus not advisable. In addition, the use of procedure will not enable to compare between coupled and non-coupled presentation forms. For example, in the investigation of coupling between modalities that involves contrasting the static-durable multimodal presentation with the dynamic-durable multimodal form.

The use of a monitoring task seems to be more appropriate: monitoring tasks require subjects to press a key on detection of a target. The description of the target can be phonetic, semantic or both. The logic is that as processing load changes across the sentence, resource allocation to the comprehension task changes as well. Thus, at the point of comprehension difficulty, many resources are devoted to comprehension, so fewer resources are available for the monitoring task. As a result, detection times are slow. An examination of the sentence comprehension should serve as a complementary task, to ensure that subjects process the sentences for comprehension.

The next chapter provides a means to relate linguistic complexity factors and multimodal presentation types in the form of predicted user cost. The MMUM attempts to characterise the cognitive structures and processes underlying human language processing in multimodal presentation. In addition, it includes an account of individual differences in verbal WM capacity. Taking all these aspects into account, the MMUM enables prediction of the effect of different multimodal presentation modes and levels of linguistic complexity on the user's ability to process information presented to both modalities.

Chapter 3

The multimodal user model (MMUM)

A MMUM is proposed, a model of multimodal language processing at the lexical, phonological, syntactic, semantic, pragmatic and propositional levels of processing. The model attempts to characterise the cognitive structures and processes underlying human language processing in multimodal presentation of text and speech. It has adopted, where possible, existing accounts of psycholinguistic processing. In this, the model has made selective use of different types of literature: reading processing literature, speech comprehension literature and a-modal language processing literature (which is non-committal about the effects of specific modalities). Specifically, the model takes an a-model language processing account (Just & Carpenter's 1992 capacity theory), elaborates it and extends it. It elaborates it by specifying the parsing system in detail (using Gibson's 1991 complexity metric), extends it downwards with an account of the modality-specific and cross-modal sub-systems, and extends it upwards with an account of the control system (Normal & Shallice's 1986 model of executive control) needed to manage this integration. As the model aims to help in the design of multimodal systems wherein users can successfully comprehend texts with a minimum of processing cost, it provides specific hypotheses (Chapter 4) regarding the variation in the (psycholinguistic) cost in response to different contents and different multimodal configurations, as specified by the MMDS. A summary of the first version of the model, as it appears in Shmueli & Dowell (1999), is provided next. This includes the structure of the model and the literature on which it is based.

3.1 The capacity theory: a model of language processing

The MMUM is based on the model of language processing postulated by the capacity theory (Just & Carpenter, 1992). The capacity theory consists of three general assumptions:

1. The comprehension of language involves both processing (symbolic manipulations) and storage of partial and final products. In this framework, capacity can be expressed as the maximum amount of activation available in WM to support these activities.
2. A common pool of resources serves both kinds of activity. Because the pool is of limited capacity, a dynamic trade-off between processing and storage is required when task demands increase. The trade-off between specific processes and storage is evident in a more global trade-off between different sub-systems. For example, when processing a syntactically complex sentence, the syntactic system requires a large amount of resources to store intermediate computational products. Less activation is then available to the cross-modal sub-systems. As a result, the computational processes in the cross-modal sub-systems are slowed down.

3. There are significant individual differences in linguistic WM capacity and these are due to variations in total capacity of resources available. These individual differences influence the points at which trade-off between processing and storage demands are necessary for a particular individual. Specifically, differences in WM capacity are mostly apparent when a linguistic task imposes an excessive load on the users.

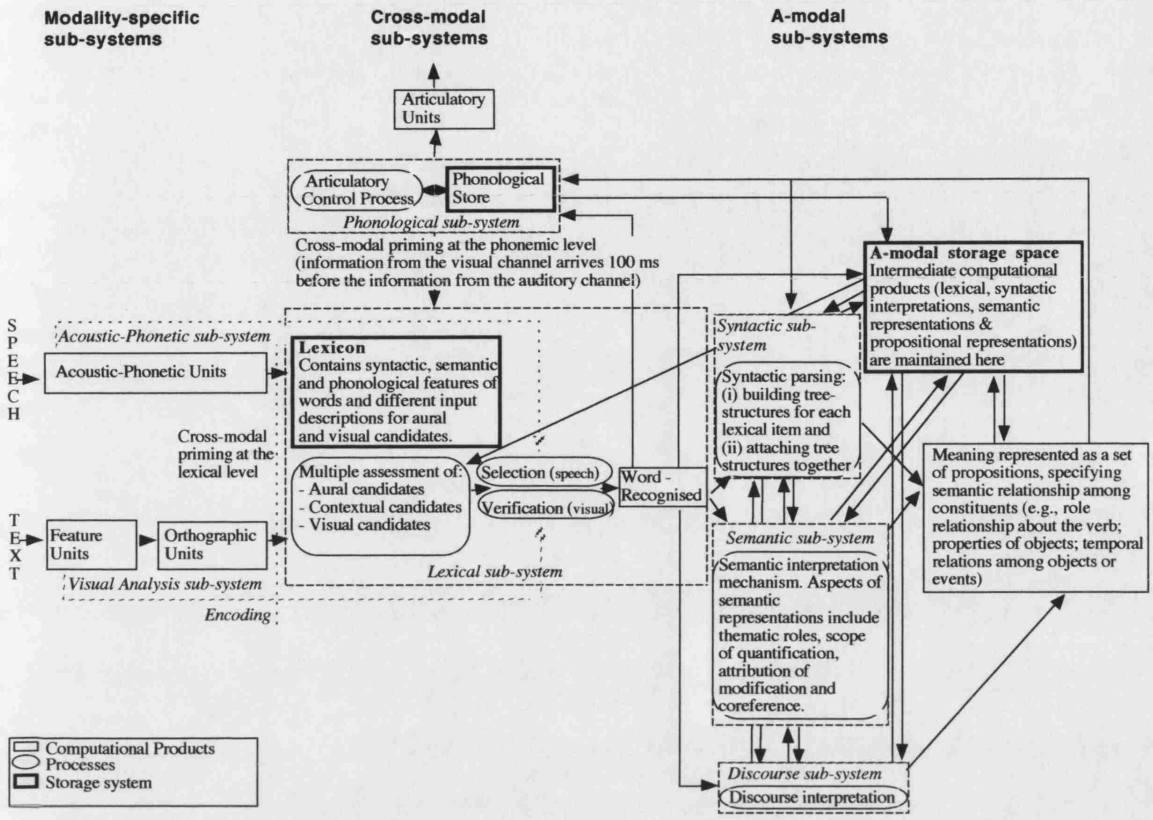
Language processing requires recognition of the lexical items represented by the surface form of language, access to their syntactic and semantic specifications, and interpretation of sentence meaning. According to the capacity theory, storage of partial and final products of these linguistic processes occurs at each level and provides input to further levels of language processing activities. Moreover, all enabled processes can execute simultaneously and generate partial products concurrently.

The capacity theory is not specific to either the processing of visual text or speech; it applies equally to both. Therefore, if one wishes to apply the capacity theory to characterise a multimodal processing of text and speech, the model must be extended to provide an account of the differentiated and integrated processes. In other words, the model must be elaborated to explain how separate concurrent processes of reading and listening become fused in a singular process of comprehension and, in particular, how control is exercised in the attempt to optimise this integration. This has several implications. The first is that structural components that do not necessarily take part in the on-line language comprehension processes as defined by the capacity theory should be incorporated to the model, if considered part of the multimodal comprehension process. Figure 3.1 presents the linguistic sub-systems that participate in the multimodal comprehension process (see sections 3.2.1-3.2.3). One of these sub-systems is the phonological sub-system that portrays Baddeley's (1986) phonological loop. This sub-system is not considered part of the language processing system as defined by Just & Carpenter (1992) but was incorporated to the MMUM, because of its role as a major contact point of cross-modal verbal input. A second implication is that the mechanisms used by the capacity theory to explain the control of language processing are not necessarily optimal for explaining the control of multimodal language processing. Norman & Shallice's (1986) model of executive control seems more appropriate for the task. The multimodal language control system will be described in section 3.2.4. Note that this control system is not represented in Figure 3.1, which is limited to the structural dimension of the model. Finally, the available capacity, the trade-offs between processing and storage and the trade-offs between the different sub-systems, all relate to the dynamics of the comprehension process and are not shown in Figure 3.1.

3.2 Structure of the model

The MMUM postulates individual and common structures, processes and the control of the processes. As demonstrated by Figure 3.1, the MMUM distinguishes between (i) modality-specific (ii) cross-modal and (iii) a-modal sub-systems that participate in the comprehension process. A description of the sub-systems involved follows.

Figure 3.1 The MMUM: Main Structures and Processes



3.2.1 Modality-specific sub-systems

The MMUM presents two modality-specific sub-systems that are tied to processing linguistic input exclusive to a specific modality. The visual analysis sub-system analyses the features of the stimulus into orthographical units, which in turn activate the lexical sub-system. Its operation is input dependent, but is also affected by a top-down activation from the lexical sub-system. The acoustic-phonetic sub-system analyses auditory input into phonetic units, and then outputs these to the lexical sub-system. The detailed structures and processes of these sub-systems are beyond the scope of this dissertation.

3.2.2 Cross-modal sub-systems

Cross-modal structures and processes refer to the lexical and the phonological sub-systems. The operation of these sub-systems is of special importance to the MMUM since this is where contact is made between spoken and written input.

The lexical sub-system

The lexical sub-system consists of activation and selection processes of word candidates and a single structure - the lexicon. The lexicon is a semantic network of lexical entries. Representations in the lexicon are assumed to correspond to functional coordinates of either the orthographical or the acoustic-phonetic input and of the syntactic, semantic and phonological specifications associated with a given lexical entry.

Human performance literature suggests that lexical-access of visual words differs from lexical-access of spoken words. Specifically, sentential context and word frequency affect different stages in the recognition process. In addition, visual word recognition is affected by word frequency and word regularity, whereas spoken word recognition is affected solely by frequency (c.f., Marslen Wilson, 1987; Paap & Noel, 1991). Despite this complexity, it is suggested that, in principle, lexical-access of both speech and visual text input has a similar form: for both modality-based inputs, there is no context driven pre-selection of candidates without some bottom-up sensory input to the lexicon. Sentential context is allowed, however, to compensate for deficiencies in the bottom up specifications of the input, by lowering the required activation threshold of the recognition process. The verification process of the visual stimulus takes place at 150 milliseconds (ms) measured from the onset of word presentation (c.f., Van Orden 1987). It consists of a top-down analysis of the visual stimulus, guided by the stored representation of the word. The selection process of the spoken word occurs at approximately 100-150 ms from the onset of word presentation. Finally, the entire word recognition process has a similar duration for both modality-based inputs: visual word recognition takes about 280 ms for isolated words[13] and 200-250 ms for words in sentence context (Rayner et al, 1982; Rayner & Pollastek, 1987). Similarly, spoken word recognition takes 300 ms for isolated words and 200 ms for words in utterance context (Marslen Wilson, 1987). Table 3.1 summarises the approximate times for the different stages of word recognition in the visual and the auditory modalities.

Table 3.1

Approximate Times for the Different Stages of Word Recognition in the Visual and the Auditory Modalities

Modality	Word verification or selection time	Total word recognition time (for isolated word)	Total word recognition time (for words in context)
Visual word	150 ms	280 ms	200-250 ms
Spoken word	100-150 ms	300 ms	200 ms

The MMUM suggests that less activation is required for lexical-access in a coupled multimodal presentation, when the visual and the auditory words are presented together at the same rate. This suggestion is supported by the cross-modal priming literature. Using various tasks (e.g., naming, lexical-decision task) and isolated word stimuli (e.g., identical and semantically related materials), it

[13] This measure reflects an artificial situation in which a reader obtains a foveal but no prior parafoveal view of a word.

has been established that both modality-based inputs make contact in the lexical sub-system and that multimodal activation of lexical representations facilitates performance. Hanson (1981) used a semantic judgement task and found that facilitation by identical words is stronger than the facilitation produced by semantically related words. However, no difference was found between the condition in which only the unattended word belonged to the semantic category of the target and the condition in which neither of the words belonged to the required category. These results indicate that name and semantic information are represented in a coding system common to the auditory and visual modalities.

Verified words access a post-lexical phonological store and an a-modal storage system that serves the syntactic and the semantic sub-systems.

The phonological sub-system

According to Baddeley (1990), the phonological loop is comprised of two components: the phonological store (a store that holds phonological, but not semantic information) and the articulatory control process (ACP). Material is registered in the phonological store either directly, through auditory presentation or indirectly when the ACP is used to convert a visual item into a phonological code. Information can be maintained in the phonological store by the process of articulation. Within the MMUM, phonological traces produced by the lexical sub-system in response to a visual input may be set up directly in the phonological store[14]. They appear to access the phonological sub-system 100 ms faster than phonological representations of speech-based input (Hanson, 1981) and but also require more activation to be maintained in the phonological sub-system by the ACP if they are to assist serial recall.

In sentence comprehension, the phonological sub-system may assist in retaining word-order information (Levy, 1978; Waters, Komoda & Arbuckle, 1985; Martin & Feher, 1990; Martin, 1990), that disappears when readers complete the syntactic analysis of the clause (Jarvella, 1971). The MMUM suggests that post-interpretative processes can make indirect use of phonological representations. These are assessed against sets of thematic roles held in a representation of the propositional content of the sentence. Reference to this form of lexical representation might initiate enough of the parse to establish the first thematic roles and other sentential features licensed by the parser, thereby also excluding those erroneous inferences made pragmatically from individual word meanings and real-world knowledge (Waters, Caplan & Hildebrandt, 1987). Most importantly, referring to phonological representations and the order of lexical items in the sentence can only reinitiate the parsing process, and cannot yield meaning directly.

[14] A suggestion made by Baddeley himself to account for the involvement of the phonological loop in homophony judgments; see Baddeley, 1986.

3.2.3 A-modal sub-systems

The MMUM suggests that processing of multimodal information converges into a single abstract representation that serves the a-modal syntactic and semantic sub-systems[15]. Once lexical entries have been located, the syntactic and the semantic features of words become available to feed into parsing and semantic interpretation processes. The products of these processes are durable sets of propositions that represent the meaning of the sentence. Propositional representations specify semantic relationship between constituents, such as the role relationships of the verb, the existence or properties of objects, and temporal relationships between objects or events (Clark & Clark, 1977; Kintsch & Van Dijk, 1978; Just & Carpenter, 1987). The quicker that these representations are derived from the input, the less dependence there will be on rapidly fading surface representations (Martin, 1990).

The syntactic and semantic sub-systems

The MMUM assumes immediate syntactic processing, an assumption common to most recent theories of language comprehension (e.g., Just & Carpenter, 1992; Marslen Wilson, 1987). Also, syntactic processing decisions make use of semantic information and pragmatic knowledge available up to that point in the sentence (Marslen Wilson & Tyler, 1987; Gibson, 1991). The MMUM similarly assumes immediate semantic processing: semantic processing is assumed to occur in parallel with syntactic processing and to make use of syntactic information, as it becomes available. Hence, syntactic and semantic sub-systems operate real-time, in parallel and establish mutually supportive dependencies.

A left-corner parsing algorithm postulated by Gibson (1991) accounts for syntactic processing in the MMUM. It consists of two stages: (i) building tree-structures (nodes) for each lexical item based on its lexical entry and grammar, and (ii) attaching tree structures together. The data structures upon which the parsing model rests are termed the buffer and the stack-set (c.f., Marcus, 1980). When a word is input to the parser, representations of each of its lexical entries are created and placed in the buffer, a one-cell data structure that holds a set of tree structures. For each of these representations, the lexical requirements, the syntactic category (and its grammar rules) and the thematic information of the structure cause a local prediction to the right for further categories. These predicted structures are called *hypothesized* nodes or H-nodes[16]. All other structures are called *confirmed* nodes or C-nodes. The stack-set contains a set of stacks of buffer cells, each of which contains tree structures for previous input.

---

[15] Martin (1990) notes that it is possible that a pre-parsing buffer could store information in the form of lexically coded phonological information (Barnard, 1985; Saffran & N. Martin, 1990), although this is not necessarily the case. A more abstract lexical code might also be employed.

[16] H-nodes are further subdivided intro optional H-nodes and obligatory H-nodes. Obligatory H-nodes are H-nodes whose presence with respect to some C-node is forced, as determined by the lexicon or a grammar rule. Optional H-nodes never obligatorily apply and are pruned from a structure at the end of the parse.

Node attachment consists of matching the structures hypothesised for the top of each stack in a stack set against nodes in the buffer. If the features of two such nodes are compatible, then an attachment takes place, the result being the unification of the stack and the buffer node. Furthermore, following the Stack Push Constraint, of those buffer structures that cannot attach to any stack structure, only H-nodes are pushed onto stacks to form a new set of stacks.

- The Stack Push Constraint (SPC): Push onto stacks only those structures whose root nodes are H-nodes and which cannot attach to any structure (for the SPC theorem, see Gibson, 1991).

This cycle continues until no words are left for attachment. If a word is input into the buffer and no attachments are possible with any of the stacks in the stack set, then processing breakdown has occurred and the input will not be processed. The products of the parsing process are maintained as propositional representations.

The parsing algorithm is constrained by memory limitations. These become apparent while processing complex sentences that require a delayed attachment of thematic roles. As described in Chapter 2, Gibson (1991) defines a set of local thematic violation constraints, each of which is associated with a numeric processing load. It is the task of these constraints (and the SPC principle) to limit the number and the kind of representations for the input string. As suggested earlier, Gibson hypothesises that there exists a maximal processing load, *K*, measurable in units (PLUs) that the human parser can maintain: the processor maintains only those structures whose processing load is less than or equal to *K* PLUs (equivalent to four local thematic violations). A structure becomes unacceptable at a particular parse state if it exceeds this processing overload constant. Gibson's account of processing load can be seen to provide a quantification of the level of activation, at the syntactic level of processing, as postulated by the capacity theory. These principles enable a complexity value to be assigned to any sentence and also to predict the exact word where processing breakdown will occur in complex sentences (see Chapter 2). Consistent with the a-modal level of processing in the MMUM, the adopted parser does not distinguish between speech and visual text input. Given a complex sentence with a processing load that is higher than *K* PLUs, the syntactic sub-system will face difficulties regardless of the input channel.

Having identified the sub-systems involved in multimodal language processing, the MMUM must include an explicit account of the sub-system that provides the executive control of these processes. Furthermore, the MMUM must include an account of the higher order reasoning processes, through which the person makes use of their task knowledge. The next section describes the different roles of the multimodal management system that controls the processing of sentences that vary in linguistic complexity and are presented in different multimodal configurations.

3.2.4 A multimodal management system

Norman and Shallice (1986) and Shallice (1988) have created a model of executive control that explains behaviour in routine and non-routine tasks. It assumes multiple sub-systems of cognitive processing. These multiple sub-systems interact to coordinate goals and actions and are controlled by

two qualitatively different mechanisms. The first level of control operates via contention scheduling (CS) which uses schemas or condition-to-action statements[17] to coordinate well-learned actions and thoughts. Once a schema is selected, it remains active until it reaches its goal or is inhibited by a competitive schema or higher-level control[18]. The CS mechanism corresponds with routine selection. When the situation is novel or highly competitive, a supervisory attentional system (SAS) intervenes and provides additional inhibition or activation to the appropriate schema for the situation. The SAS has access to the overall representation of the environment, the current goals of the person and the cognitive capacities available to support these goals. This is in contrast to the CS mechanism that only controls lower level competition among schemas.

This model of executive control seems a highly appropriate candidate for a multimodal language control-system. The MMUM assumes that, within this control system, schemas are specialised for a particular control of language processing and for the integration between modalities. The processing of long-complex sentences presented to both modalities qualifies as an example of a non-routine task, where dividing attention between modalities might fail to occur. On the other hand, the processing of short-simple sentences can qualify as an example of a routine task, where the integration between modalities can be achieved by means of the CS. In order to simplify the suggested account, it is assumed that the SAS supervises the coordination of processing between words that are seen and words that are heard. Specifically, it is proposed that this coordination is achieved by means of *synchronisation* so as to maximise the multimodal activation of cross-modal sub-systems by redundant input. A primary regulator of this synchronisation is the rate at which a visual item is brought to foveal vision when the visual text has a static-durable form.

As suggested earlier, name and semantic information are automatically activated in the lexical sub-system by redundant input of single words. Processing *sentences* presented to both modalities raises a synchronised processing problem when the visual text has a static-durable form: whereas for early sentential components, the lexical sub-system is accessed with redundant information, late sentential components will not necessarily make contact on a redundant multimodal basis without additional control of reading pace. This is because readers do not necessarily process a statically presented text in a linear fashion, They can skim the text to locate particular information and also employ regressive eye-movements to earlier parts of the text when complexity arises (see Chapter 2, section, 2.3.1). The SAS is therefore required to control the synchronisation between reading pace and listening to enable this multimodal activation while processing sentential constructions (i.e., to bring "out of phase" stimuli into phase during processing).

---

[17] This research uses the concept of 'schema' to express the control of processing. The term is not used to express the user's mental model of the world (e.g., to the pragmatic knowledge that dogs bite men.

[18] Some significant differences are evident in the assumptions that underlie the Norman & Shallice model of executive control and the Just & Carpenter capacity theory. These differences have yet to be demonstrated computationally. For the purposes of the MMUM, the similarities are sufficient to justify adoption of the model of executive control.

The MMUM also assumes that language processing demands (processing and storage) share resources with the SAS. The SAS monitors the performance of the language processing system and controls the competition between different sub-systems for activation. The control is exhibited by modifying the activation of particular cognitive operations concerning language processing (e.g., allocation of resources to storage of intermediate representations in syntactically complex sentences). The SAS ensures that any specific cognitive operation does not capture resources required by other cognitive operations for sentence comprehension. Finally, the shared resources assumption implies that increasing sentence length and complexity imposes demands not only on the resources that are used by the language processing system, but also on the resources used by the SAS. Given that linguistic task-requirements exceed the available capacity, the SAS will fail to supervise the synchronisation between modalities.

According to this framework, at the very general level, a coupled multimodal display of visual text and speech enables the user to divide attention between modalities throughout sentence computation. When the visual and the auditory words are presented simultaneously at the same rate, more resources can be allocated to the language system. The multimodal contact in the lexical sub-system enables processing to feed the a-modal sub-systems with consistent representations of the verified words. Given the assumption that a common pool of resources serves the language processing system, this implies that more resources are available to the syntactic and the semantic sub-systems. Furthermore, consistent representations maintained by the phonological sub-system enable post-interpretative use of phonological information. This does not imply, however, that sentence processing can always be achieved by means of the CS when presentation consists of coupled visual and auditory words. Short-simple sentences presented to both modalities in a coupled manner can be processed by means of the CS. On the other hand, long-complex sentences require the intervention of the SAS regardless of the coordination between modalities.

3.3 Conclusions

The MMUM is more than a literature review. This is a cognitive model that characterises the structures and processes underlying multimodal language processing, including the supervisory attentional mechanisms that coordinate the processing of language in parallel modalities. The originality of the model is in the coherent integration of different literatures so that the model is able to make predictions about the cognitive effects of particular presentation configurations. The MMUM can take as input specifications of the content and its representational form (as defined by the MMDS) and return as output a statement of the expected user cost. The next chapter combines the two models. It provides detailed predictions of the MMUM about levels of user cost for different contents that are presented in different configurations. The chapter details possible effects of media realisation and coordination on user cost and predicts how capacity limitations of the user affect processing cost in multimodal language processing. The subsequent chapters report a set of studies in which the hypotheses were tested, the design of those studies was therefore closely determined by the MMUM.

Chapter 4

Predictions and their investigation

The MMUM characterises the cognitive processes of reading and listening to single sentences and how these processes may combine. The MMDS characterises the fundamental dimensions of multimodal user interfaces. Here, these two theoretical structures are brought together to make specific predictions about levels of cognitive cost which users will experience with different multimodal interface designs. These predictions form the basis of the subsequent studies reported in the thesis.

4.1 Using the user model and the design space to predict the general effects of media realisation and coordination on user cost

The MMDS specifies two properties that distinguish between different media realisation techniques: the durability and the dynamism of the selected media. This section uses the MMUM to make general predictions about the combined effect of these two properties on the user's cognitive cost.

The MMUM proposes that dynamic-transient media, dynamic-durable visual text and static-durable visual text do not involve the same memory demands. These demands do not refer to the syntactic complexity constraints specified by Gibson. They refer to the storage and computational demands imposed by different presentation techniques for a successful processing of a given sentence. Significantly, the model identifies the *durability* of the visual text as the principal determinant of memory demands, as now discussed.

The MMUM describes how readers may make visual regressions to earlier portions of a sentence to retrieve lost information when trying to compute the sentence. It is suggested that a *static-durable* visual presentation encourages users to allocate resources to immediate sentence computation rather than to storage of intermediate computational products. This repeated de-allocation of resources may induce a kind of "forgetting" by gradually reducing the amount of resources necessary to keep various intermediate and/or final products of comprehension active in WM. According to Miyake, Carpenter & Just (1994), if the "forgetting" is so extensive that a necessary piece of earlier information is not available in WM when it is needed, the parser may not be able to compute new elements critical for the comprehension of the sentence. Even if the critical information is available at a given point in time, the products of the computation may be lost by the time the end of a sentence is reached. That is, unless the required information is collected again using regressive eye-movements. According to this paradigm, the information collected by regressive eye-movements can help in the interpretation of long-simple sentences. For long-complex sentences, where storage and computation

demands exceed the available capacity, this will not prove useful. Consistent with Gibson's metric, the MMUM assumes that the syntactic parser will not have sufficient resources available for the delayed assignment of thematic roles required for the comprehension of complex sentences.

In contrast, a *dynamic-transient* media (e.g., speech, dynamic-transient visual text) imposes high memory demands on the users when the sentences presented are relatively long or complex. This is because verbal information presented by a dynamic-transient media must be processed in real time; if processing of the information cannot be completed immediately, the user must maintain their primary representations for later processing.

The status of a *dynamic-durable* visual presentation is different: a dynamic-durable visual text forces the pace of reading but nevertheless provides durability. Theoretically, this presentation technique enables the use of regressive eye-movements. However, one should note that since dynamic media attracts attention automatically, regressive eye-movements are more difficult to carry out under dynamic-durable conditions. Moreover, given a fast rate of presentation, regressive eye-movements must be conducted at the expense of processing of a later sentential component.

Media coordination depends on the realisation of the visual text in one of the above three forms. The MMUM identifies the coordination by synchronisation between the visual and the spoken words as the main predictor of user cost in multimodal processing. A visual text will only be recognised as synchronous with speech, when that text is 'served' to the reader dynamically (i.e., word by word). When the visual text is presented in a *static-durable* form, the addition of speech results in a non-coupled multimodal presentation. The users can determine the synchrony in such a non-coupled presentation through coordinating their reading with their listening. This presentation form has the advantages of allowing the user to scan and to skim the visual text, including making regressive eye-movements to previously processed portions of text. However, despite its high level of visual-processing control, the MMUM specifies a synchronised processing problem for this form of presentation: whereas for early sentential components, the lexical sub-system is accessed with redundant information, late sentential components will not necessarily make contact on a redundant multimodal basis due to resource constraints. Although the model assumes that the SAS is capable of supervising the coordination of processing of visual and auditory stimuli by bringing "out of phase" stimuli into phase, it also assumes that increasing processing demands will impair its capability to supervise this synchronous processing. Furthermore, the use of regressive eye-movements, associated with increasing processing demands, will impair processing. Specifically, the incompatibility of the spoken information with the recollected visual information will produce an interference effect at the cross-modal sub-systems for all users. As a result, conflicting representations will access the a-modal storage space that serves the syntactic and semantic sub-systems. Significantly, the MMUM assumes that, with increasing processing demands, the SAS will not have sufficient resources available to supervise the competition for activation between the language sub-systems and to oversee the coordination of information between modalities.

In contrast, both *dynamic* multimodal formats avoid the problem of supervising synchronous processing. According to the MMUM, the dynamic formats may form a succession of cross-modal priming at the word level as the sentence is processed. Specifically, when the visual and the auditory stimuli are presented together at the same rate, more resources can be allocated to the language system. The multimodal contact in the lexical sub-system enables processing to feed the a-modal sub-systems with consistent representations of the verified words. Given the assumption that a common pool of resources serves all language processing sub-systems, this multimodal activation of the cross-modal sub-systems implies that more resources are available to the syntactic and the semantic sub-systems. Furthermore, consistent representations maintained by the phonological sub-system enable post-interpretative utilisation of phonological information.

Finally, as suggested earlier, the two dynamic multimodal formats differ in terms of the durability of the visual presentation; whereas information presented by dynamic-transient multimodal form must be processed in real time, the dynamic-durable multimodal presentation supports a (limited) use of eye regressions. It is suggested that under increasing processing demands, the user may attempt to use eye regressions to recollect early sentential components while attending to the spoken continuation of the sentence. The incompatibility of the spoken information with the visual information might cause some interference in processing. However, the synchronous processing is not expected to breakdown because it can easily be restored through re-focusing attention on the "leading edge" of the words accumulating on the visual display.

Having predicted the general effects media realisation and coordination have on user cost, the MMUM provides specific hypotheses about the effect of different levels of linguistic complexity on user cost for different multimodal configurations. These are presented in the next section.

4.2 Predicting the effect of linguistic complexity on user cost for different multimodal configurations

The following table summarises the predicted cost of processing sentences that vary in their linguistic complexity when presented in different multimodal configurations.

Table 4.1

Predicted User Cost for Different Multimodal Configurations: By Linguistic Complexity

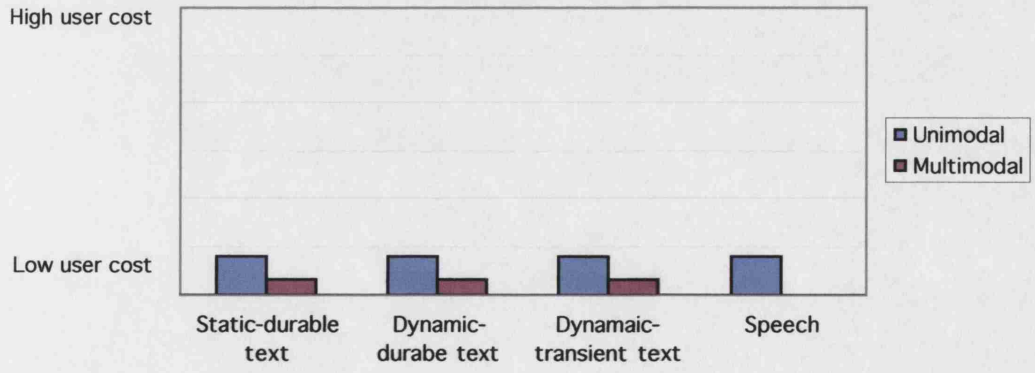
	Multimodal configurations		
Linguistic complexity	Static-durable	Dynamic-transient	Dynamic-durable
Short-simple	facilitation	facilitation	facilitation
Long-simple	slight interference	slight facilitation	slight facilitation
Long-complex	interference	similar (poor) performance	possible interference

4.2.1 Short-simple sentences

Regardless of their pragmatic complexity value, short-simple sentences involve an immediate assignment of thematic roles. The model predicts a superior performance for short-simple sentences presented to both modalities than for a unimodal presentation of such sentences. Given a static-durable multimodal presentation of short-simple sentences, the coordination between visual text and speech displays can be achieved by the CS. Similarly, regardless of the durability of the visual presentation, the CS can maintain a synchronous processing of short-simple sentences that are presented dynamically to both modalities (regressive eye-movements are assumed to be of minor importance when the sentences presented to both modalities are short and simple). For this synchronous processing, the multimodal activation of the lexical-access system will optimise information processing in the a-modal sub-systems that are fed by consistent representations of the verified words. Following the assumption of a common pool of resources shared by the language processing sub-systems, a multimodal activation of the cross-modal sub-systems also implies that more resources are available to syntactic and semantic sub-systems. The syntactic sub-system is not assumed to face difficulties, since thematic roles are immediately assigned in simple sentences. The semantic relationships between constituents are easily derived. Finally, the consistent representations maintained by the phonological sub-system enable post-interpretative utilisation of phonological information, if required.

The predicted user cost for short-simple sentences in various presentation configurations can be seen in Figure 4.1

Figure 4.1
Predicted User Cost as a function of Presentation Type for Short-Simple Sentences

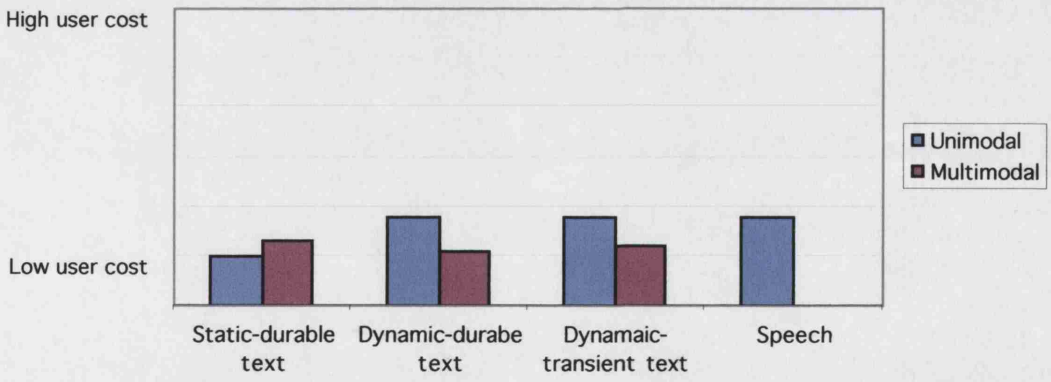


4.2.2 Long-simple sentences

For the medium linguistic demands of long-simple sentences, the MMUM predicts an asymmetric relationship between the processing of static-durable visual text and the processing of speech: text will facilitate speech, but speech will not facilitate text. Specifically, a static-durable multimodal presentation of long-simple sentences is expected to reduce user cost in comparison to a dynamic-transient spoken presentation of such sentences (see Figure 4.2). Since the speech must be processed as it arrives, displaying the redundant visual message in a static-durable form would enable further retrieval of intermediate representations. On the other hand, the MMUM predicts a slightly better performance for a static-durable visual presentation than for a static-durable multimodal presentation of long-simple sentences. For the multimodal presentation, the model assumes that the length factor is sufficient to slightly hinder the coordination of information between modalities. In addition, regressive eye-movements are expected to produce a temporary interference between the visual and the auditory representations at the cross-modal systems. As a result, conflicting representations will access the a-modal storage space that serves the syntactic and semantic sub-systems. However, since long-simple sentences involve an immediate assignment of thematic roles and the semantic relationships between constituents are easily derived, this interference due to regressive eye-movements is not expected to impair performance significantly[19]. Multimodality is not expected to improve performance either, since a solely visual presentation of static-durable text enables the same regressive eye-movements without the interference of speech.

Figure 4.2

Predicted User Cost as a function of Presentation Type for Long-Simple Sentences



As noted earlier, dynamic-transient media creates high memory demands. A multimodal presentation of long-simple sentences in a dynamic-transient form is expected to slightly reduce user cost. According to the MMUM, when the visual and the auditory stimuli are presented together at the same rate, more resources can be allocated to the language system. The multimodal contact in the lexical sub-system enables processing to feed the a-modal sub-systems with consistent representations of the verified words. Moreover, given the assumption that a common pool of resources serves all language

[19] See however the differential predictions for low and high span subjects, outlined in section 4.3.

processing sub-systems, this multimodal activation implies that more resources are available to the syntactic and the semantic sub-systems. The syntactic sub-system is not assumed to face difficulties, since thematic roles are immediately assigned. The semantic relationships between constituents are easily derived. Finally, the consistent representations maintained by the phonological sub-system enable post-interpretative utilisation of phonological information. Referring to consistent phonological representations such as these, may compensate for the absence of regressive eye-movements in the dynamic-transient presentation condition.

Similarly, long-simple sentences presented in a dynamic-durable multimodal form should yield a slightly better performance than a dynamic-durable presentation to the visual modality only. As described earlier, the dynamic-durable multimodal presentation enables the user to use eye regressions to recollect early sentential components while attending to the spoken continuation of the sentence, although the machine-governed reading pace makes them more difficult to carry out. The incompatibility of the spoken information with the visual information might produce a slight interference effect. However, the synchronous processing is not expected to breakdown because it can be easily restored through re-focusing attention on the "leading edge" of the words appearing on the visual display. Under these circumstances, the SAS is expected to successfully accommodate the coordination of information between modalities on one hand, and the demands of the a-modal sub-systems for activation on the other.

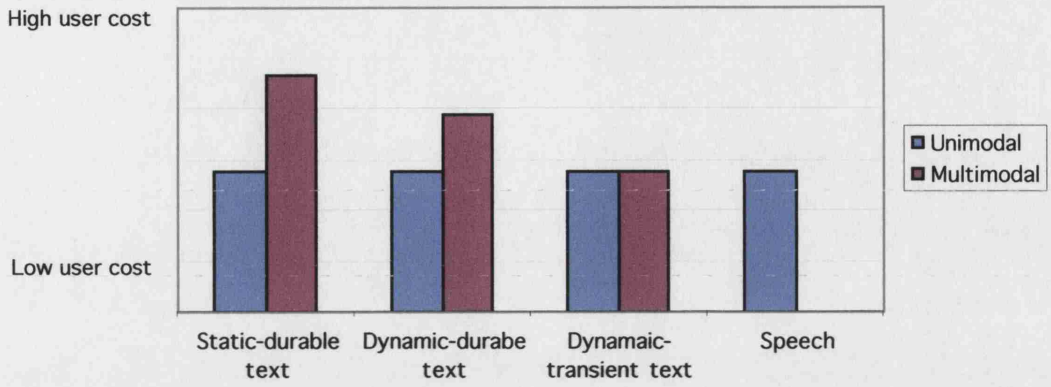
4.2.3 Long-complex sentences

Gibson (1991) claims that for complex sentences with a processing load that is higher than *K* PLUs, syntactic parsing will fail regardless of the selected mode of presentation. Therefore, users will have to operate in a "repair" mode, if they are to comprehend the sentence at all. It is against this background that the question is raised whether a multimodal presentation might have a beneficial effect on user cost and whether, at base, it might prevent the parse from failing.

The MMUM proposes that this is not the case. The model predicts an inferior performance for long-complex sentences in a static-durable multimodal presentation, compared with either static-durable visual text or speech presentation of such sentences (see Figure 4.3).

Figure 4.3

Predicted User Cost as a function of Presentation Type for Long-Complex Sentences



The excessive storage demands imposed by the need to maintain several unassigned thematic roles in memory impose high demands on the syntactic sub-system (c.f., Gibson, 1991). The model also suggests that the semantic sub-system faces difficulties in specifying the semantic relationships between constituents. Trying to address these requirements, the SAS is assumed to allocate more resources to these demanding sub-systems at the cost of less activation available to the lexical and phonological sub-systems. As a result, the computational processes in the cross-modal sub-systems are expected to slow down.

Furthermore, regressive eye-movements are not expected to assist the interpretation of long-complex sentences in the static-durable multimodal format. Even if the durability of the visual text encourages users to allocate resources to sentence computation rather than to storage of intermediate computational products, this repeated de-allocation of resources may induce "forgetting" of intermediate products of computation. The unavailability of an earlier piece of information will not allow the parser to compute new elements necessary for the comprehension of the sentence. Consistent with Gibson's metric, the MMUM predicts that the syntactic parser will not have sufficient resources available for the delayed assignment of thematic roles required for the comprehension of complex sentences.

The shortage of resources for activation of linguistic operations is also assumed to impair the multimodal synchronisation function supervised by the SAS, since both functions draw upon the same limited pool of resources. Given that the SAS fails to control the coordination of information between modalities, both cross-modal sub-systems will process conflicting representations arriving from the two information channels. The fusion suggested at an a-modal level of processing implies that conflicting representations will access the a-modal storage space that serves the syntactic and semantic sub-systems.

Finally, the additional speech channel enables the user to use eye regressions to recollect early sentential components while attending to the spoken continuation of the sentence. Theoretically, this

strategy enables a switch of attention between modalities, so as to assign thematic roles across them. However, it is not expected to alleviate the high processing cost of long-complex sentences. Rather, the MMUM assumes that the SAS will not have sufficient resources available for its supervision of (i) the competition between the language sub-systems for activation and (ii) the assignment of thematic roles across modalities.

In contrast, the dynamic-transient multimodal technique does not enable the user to perform regressive eye-movements and is, therefore, more resistant to any violation of synchronisation. Multimodality is not expected to improve performance either since the simultaneous cross-modal activation of the lexical-access sub-system will be negligible under excessive complexity conditions.

Finally, the dynamic-durable multimodal presentation enables a limited use of eye regressions while attending to the spoken continuation of the sentence. The excessive processing demands of long-complex sentences suggest however that this strategy will not be successful. Specifically, the incompatibility of the spoken information with the recollected visual information will produce an interference effect in the cross-modal sub-systems. As a result, conflicting representations will access the a-modal storage space that serves the syntactic and semantic sub-systems. The MMUM assumes that the SAS will not have sufficient resources available to supervise the competition for resources between the language sub-systems and the coordination of information between modalities. It is, therefore, unlikely that regressive eye-movements will support the cross-modal assignment of thematic roles and, hence, the comprehension of long-complex sentences in a multimodal presentation.

4.3 Predicting the effects of Individual differences

A central claim of the MMUM is that the verbal WM capacity of the user affects user cost when processing sentences presented to both ear and eye. This claim relies on the account of individual differences posited by the capacity theory. The account is outlined in the beginning of Chapter 3: there are significant individual differences in linguistic WM capacity and these are due to variations in total capacity of resources available. These individual differences influence the points at which a trade-off between processing and storage demands are necessary for a particular individual during sentence processing. In multimodal sentence processing, the size of verbal WM capacity may affect the extent to which, for a given level of load, the SAS can successfully accommodate both the competition between the language sub-systems for activation and the coordination of information between modalities. If this claim proves correct, then individual differences in verbal WM capacity will be of major significance in the design of multimodal applications intended for the elderly as, according to Just and Carpenter (1992), reduction in WM capacity occurs with aging. This of course also means that individual differences in verbal WM capacity will be of relevance in the design of multimodal applications intended for the general population since all types of users need to be accounted for.

A task was devised by Daneman & Carpenter (1980) to assess verbal WM capacity. The task draws simultaneously on the processing and storage resources of WM. It requires subjects to read aloud a series of sentences (of approximately 15 words each) and then recall all their final words. Users can be classified according to their performance as being *High* span, *Medium* span, or *Low* span. The capacity theory suggests that the comprehension processes used in reading the test sentences consume less of the WM resources of high span readers. Therefore, high span readers have more resources left to retain the final words of the sentences.

According to the capacity theory, individual differences in verbal WM capacity are mostly apparent when a linguistic task imposes an excessive load on users. On this basis, the MMUM hypothesises that individual differences in verbal WM capacity will affect performance when the combination of the multimodal presentation technique and the linguistic complexity of the presented materials imposes a high processing load on the users. The higher the processing load, the more apparent the difference will be[20].

According to the MMUM, the SAS draws upon the same pool of resources as the language system for its linguistic supervision functions. The efficiency of the SAS depends on the available verbal WM capacity for its linguistic operations and is assumed to be lower for low span users. So, the SAS of low span users will face more difficulties than that of high span users under lower processing demands and this difference is expected to be significantly larger for high processing demands. Thus, the model forecasts both groups of users to show similar performance rates for (low load) short-simple sentences presented to both modalities. The difference between the two capacity groups is expected, however, to influence performance when processing long-simple sentences presented to both modalities (see Figures 4.4 and 4.5).

---

[20] Note that Waters & Caplan (1996a) suggest an alternative to the capacity theory of Just & Carpenter (1992). They claim that there are specialisations within the verbal processing resource system for different verbally mediated tasks. Their working hypothesis is that one resource pool is used in obligatory, on-line psycholinguistic operations in the comprehension process and another in controlled, verbally mediated tasks (Caplan & Waters, 1990; Waters, Caplan & Rochon, 1995; Waters & Caplan, 1996b). For example, in spoken language processing the obligatory on-line psycholinguistic operations in the comprehension process consist of those operations that transform the acoustic signal into a preferred, discourse-coherent, semantic representation. These include syntactic parsing as well as acoustic-phonetic conversion, lexical-access, assignment of intonational contours, determination of sentential semantic values such as thematic roles, and determination of discourse-level semantic values such as topic and coherent co-reference. The domain of on-line language comprehension contrasts with conscious, controlled, and verbally mediated processes, such as the deliberate search through semantic memory for a piece of information, explicit reasoning, and other tasks, for which Waters & Caplan postulate a different set of resources. According to these researchers, the span task taps resources dedicated to these controlled and verbally mediated processes. Therefore, individual differences in verbal working-memory capacity are not assumed to affect processing of sentences that vary in linguistic complexity. The model assumes the validity of the capacity theory, rather than the alternative proposed by Waters & Caplan.

Figure 4.4

Low Span Subjects: Predicted User Cost as a function of Presentation Type for Long-Simple Sentences

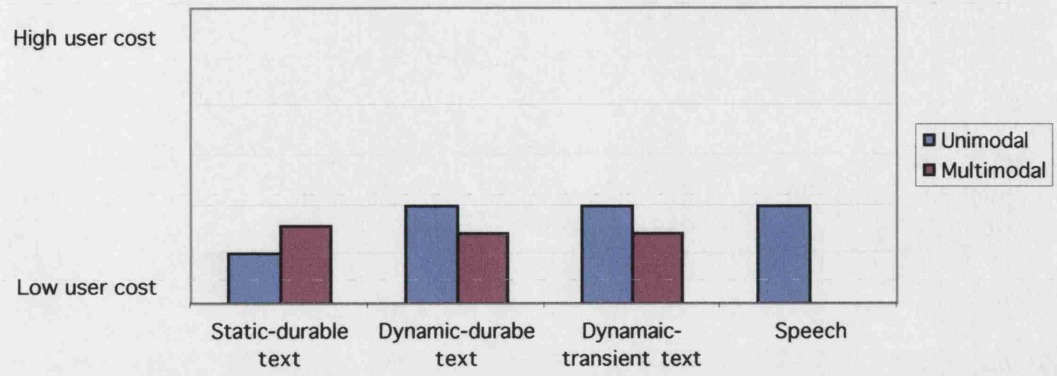
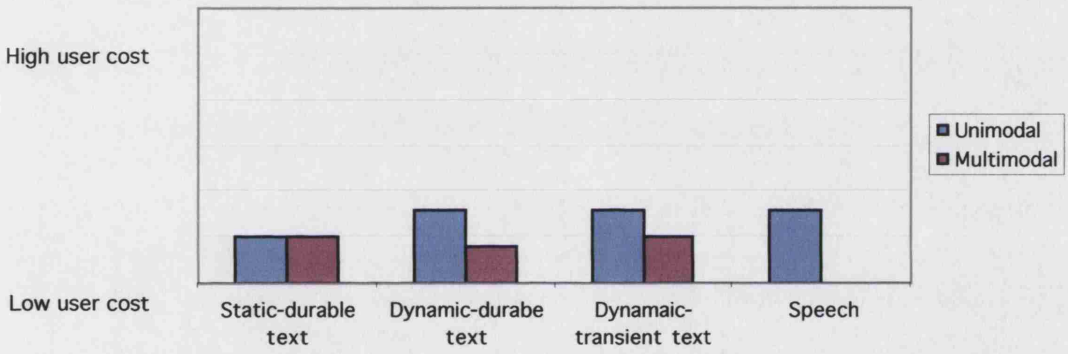


Figure 4.5

High Span Subjects: Predicted User Cost as a function of Presentation Type for Long-Simple Sentences



With the exception of the static-durable visual presentation that enables all users to flexibly recollect earlier portions of the sentence, low span users are expected to perform worse than high span users in all presentation conditions. Moreover, multimodality is expected to affect their performance in a differential manner, depending on the properties of the visual text. This differential effect will be most apparent when the visual text is presented in a static-durable form; the model assumes that low span users rely greatly on the durability of the visual text while processing long-simple sentences. Both (static and dynamic) durable multimodal presentation conditions enable the users to attend to the spoken continuation of the sentence while performing eye regressions. The incompatibility of the spoken information with the recollected visual information is expected to produce an interference effect for the low span users. Their SAS is assumed to have insufficient resources to accommodate both storage and computational demands of long-simple sentences and the coordination of information between modalities. However, this interference will be manifested only for the static-durable multimodal presentation. As suggested earlier, the machine-governed reading pace in the dynamic-durable multimodal condition makes it more difficult to perform regressive eye-movements and hence decreases the beneficial effect of the durable presentation. Yet, the leading edge of the

visual text in this presentation format is assumed to make synchronous processing easier to restore. As a result, performance of all users in this condition is expected to be slightly better than their performance in the dynamic-durable visual condition. In contrast, the static-durable multimodal presentation is expected to impair performance of low span users compared to the static-durable visual condition. Different predictions relate to high span users who are not expected to experience a significant interference in the static-durable multimodal condition. These users are assumed to have sufficient resources to accommodate both storage and computational demands of long-simple sentences and the coordination of processing between modalities.

A multimodal presentation of long-complex sentences is expected to yield a different pattern of performance (see Figures 4.6 and 4.7). Here, low span users are expected to perform worse than high span users in all presentation conditions. Moreover, for both (static and dynamic) durable multimodal conditions, performance of all users will be lower than their performance in the corresponding (solely) visual conditions. The multimodal interference effect for low span users is expected to be significantly stronger than for high span users. The reduced capacity of low span users suggests that their SAS will be less successful in supervising (i) the competition for resources between the language sub-systems and (ii) the coordination of processing between modalities. Finally, the interference is expected to be stronger in the static-durable than in the dynamic-durable multimodal condition, as synchronised processing is assumed to be easier to restore in the latter condition.

Figure 4.6

Low Span Subjects: Predicted User Cost as a function of Presentation Type for Long-Complex Sentences

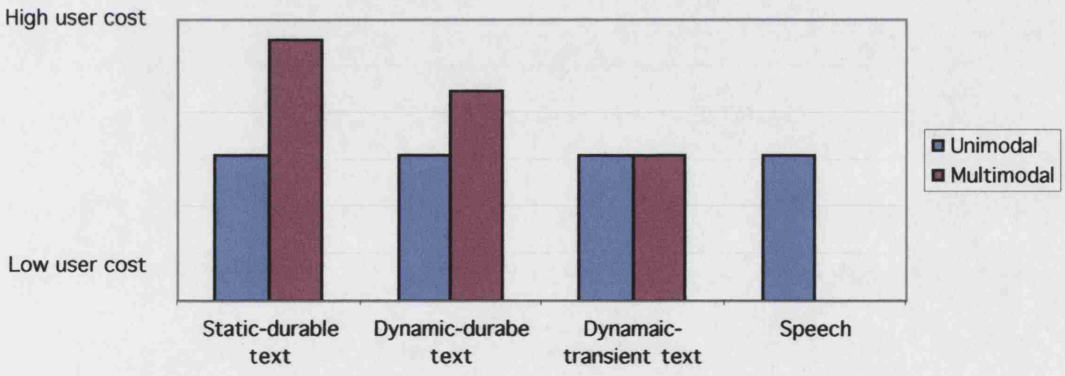
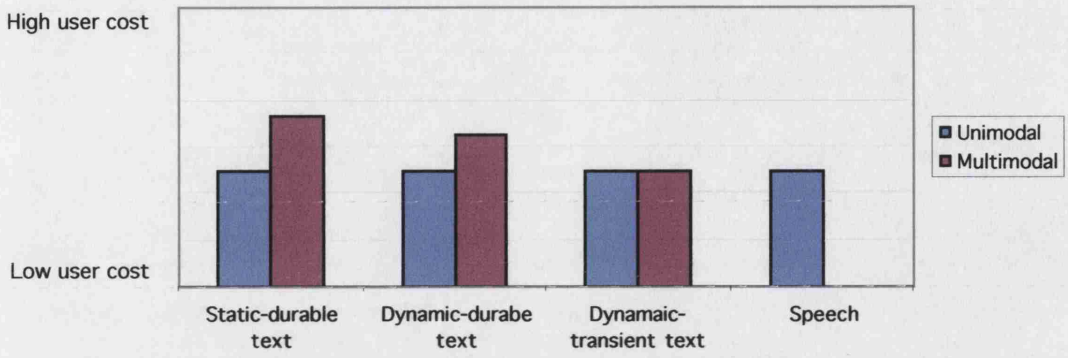


Figure 4.7

High Span Subjects: Predicted User Cost as a function of Presentation Type for Long-Complex Sentences



4.4 Conclusions

The following chapters in this thesis report a series of experimental studies, which aimed to test predictions arising from the MMUM. In this way, the studies provide an empirical validation and refinement of the MMUM, and they also support the derivation from the MMUM of guidelines for multimodal user interface design. The studies involve a systematic manipulation of syntactic complexity (and to some extent, a systematic manipulation of sentence length) using various configurations of multimodal presentation. Specifically the studies address the following questions:

1. Assuming a transition from a facilitatory synchronous processing to an interfering asynchronous processing in a static-durable multimodal presentation of long-simple and long-complex sentences, what is the effect of the attended *modality* on user cost? (In other words, in processing the early parts of a sentence, when load is low, does the visual text facilitate speech processing or does speech facilitate visual text processing? Also, in processing the later parts, when the load is medium or high, does the visual text impair speech processing or does speech impair visual text processing?)
2. What is the effect of *multimodality* in itself (i.e., the pure effect of combining modalities independently of the representational properties of the modalities) on user cost, given variation in syntactic complexity?
3. In a dynamic presentation, what is the effect of the *durability* of the visual text on user cost, given variations in syntactic complexity and multimodality?
4. In a durable presentation, what is the effect of the *dynamism* of the visual text on user cost, given variations in syntactic complexity and multimodality?
5. What is the effect of users' *individual differences* in verbal WM capacity on user cost, given variations in syntactic complexity, multimodality, visual durability and visual dynamism?

The experiments described in this dissertation all make a systematic variation of syntactic complexity based on Gibson's (1991) complexity metric. The first experiment uses four presentation conditions

to investigate the transition from a facilitatory synchronous processing to an interfering asynchronous processing in a static-durable multimodal presentation of long simple and complex sentences. It compares static-durable visual text with a (static-durable) multimodal visual-attend presentation and speech presentation with a (static-durable) multimodal auditory-attend presentation to attribute the source of facilitation or interference in the processing to specfic modalities. To manipulate attention, the experiment uses a word monitoring-task with target words presented in one modality. By placing the targets in either early or late positions in the test sentences, the experiment aims to examine how user's cognitive cost changes through the course of processing a sentence and also, to (indirectly) assess the role of the length factor in multimodal presentation.

The second experiment uses the dynamic-transient and the dynamic-durable presentations to assess the effect of the durability of a dynamic visual text on comprehension of long simple and complex sentences in both unimodal and multimodal presentations. Also, the addition of speech to the dynamic-transient visual condition is used to assess the effect of multimodality in itself on user cost, given variations in syntactic complexity.

The third experiment examines the combined role of coupling and dynamism in comprehending long simple and complex sentences presented to both modalities. Two durable presentation techniques are used: dynamic-durable and static-durable visual text. Since both of these techniques enable the user to regress back to previously read portions of a sentence, the role of dynamism in processing visual sentences can be assessed. Finally, in the dynamic-durable multimodal condition the visual and the spoken words are presented together at the same rate, whereas in the static-durable multimodal condition presentation is inherently non-coupled. This variation of coupling redundant visual and auditory words in the two multimodal conditions enables a study of the combined role of dynamism and coupling in multimodal presentation.

The results of these studies indicate the validity of some features of the model but also the need for further refinement. In giving an account and discussion of the studies, each of the following chapters also identifies the required changes.

Chapter 5

Experiment 1: dynamic variation of user cost and the role of the attended modality

This chapter reports an exploratory study, demonstrating an empirical intersection between the MMUM and the MMDS. Specifically, the study attempted to validate predictions derived from the MMUM concerning the effect on user cost of focussing attention on one modality in a multimodal presentation, given variations in syntactic complexity. Sentence presentation had four formats: visual text only, multimodal visual-attend, auditory only and multimodal auditory-attend. All visual texts were static-durable. Syntactic complexity was systematically varied based on Gibson's (1991) complexity metric. To manipulate attention, the experiment used a word-category monitoring-task with target words presented in one modality. By placing the target words in either early or late positions in the test sentences, the experiment also sought to examine how user's cognitive cost changes through the course of processing a sentence (i.e., as linguistic demands accumulate). As well as monitoring for the target words, subjects had to perform a comprehension task for each sentence. This dual-task paradigm enabled the investigation of both word-recognition local processes and interpretative processes underlying multimodal processing. Contrary to the predictions of the MMUM, the word-monitoring times indicate that subjects do not synchronise between the auditory and the visual modalities in a static-durable multimodal presentation. Furthermore, the comprehension measure indicates that a static-durable multimodal presentation of complex sentences is superior to a unimodal presentation of such sentences. Two alternative explanations are provided to account for this result. According to substantive explanation, subjects are able to advantageously switch attention between modalities so as to perform a delayed assignment of thematic roles across them. The alternative methodological explanation proposes that the observed facilitation is an artifact caused by qualitative variations in the monitoring task. These variations can be controlled in further experiments.

5.1 Introduction

There have been many studies showing that information presented to a modality other than the one to which the subject is currently responding can facilitate or hinder processing of an attended stimulus. The cross-modal priming studies conducted by Greenwald (1970), Lewis (1972) and Kirsner & Smith (1974) mentioned in Chapter 1 provide only limited examples. These studies aimed to examine

whether the source of the facilitation or interference in cross-modal priming is at the lexical or the phonological levels of processing. As these examples indicate, the available knowledge about multimodal processing is mostly limited to the single word-level. It is yet to be determined whether users divide or switch attention between modalities during multimodal sentence processing. Manipulation of attention is therefore required in order to learn about multimodal processing of whole sentences.

5.1.1 Selecting the presentation configurations

Experiment 1 included two modality-based conditions: visual-based and auditory-based. In the visual-based conditions, a static-durable visual text was compared with a static-durable multimodal presentation, in which subjects were asked to attend the visual channel and to ignore the concurrent auditory channel. In the auditory-based conditions, spoken sentences were compared with a static-durable multimodal presentation, in which subjects were asked to attend to the auditory message and to ignore the static-durable visual text. This manipulation of attention was necessary to interpret the source of facilitation or interference in each of the multimodal conditions. As suggested earlier, during the processing of a single sentence there may occur a transition from a facilitatory synchronous processing, when the user is able to coordinate the processing of multiple modalities, to an interfering asynchronous processing, when this coordination fails. The transition is due to the accumulation of linguistic demands within a sentence. Manipulating attention can demonstrate whether visual text facilitates or impairs speech processing, or the converse - whether speech facilitates or impairs visual text processing – at different points during the processing of a sentence, as load rises.

Table 5.1 presents the four presentation conditions in this experiment:

Table 5.1
Presentation Conditions in Experiment 1: By Modality and Multimodality

Modality	Multimodality	Presentation conditions
Visual-based conditions	Unimodal	static-durable visual text
	Multimodal	(static-durable) multimodal visual-attend (MMVA)
Auditory-based condition	Unimodal	speech
	Multimodal	(static-durable) multimodal auditory-attend (MMAA)

5.1.2 Selecting the sentences

Two sentence structures were selected to assess the effects of syntactic complexity on the user's ability to coordinate between visual and auditory verbal information:

- Right-branching (e.g., The dog bit the woman that likes the man that eats red meat).
- Doubly-embedded (e.g., The man that the woman that the dog bit likes eats red meat).

These syntactic structures are widely used in psycholinguistic studies, due to their qualities of imposing differential demands on the parser while preserving the semantic relationships between the nouns and the verbs (e.g., Eady & Fodor, 1981; Frazier, 1985). As suggested in Chapter 2, doubly-embedded sentences impose an excessive processing load on the user: $5X_{INT}$ processing load units (PLUs) (or five local thematic violations), as specified by Gibson's (1991) complexity metric. Again, an acceptable processing load is defined by Gibson as $K \leq 4X_{INT}$ PLUs. Right-branching sentences impose an acceptable processing load on the user (X_{INT} PLUs), by allowing an immediate assignment of thematic roles. The processing breakdown point in each of the complex doubly-embedded sentences was determined using Gibson's (1991) LRNH principle, outlined in Chapter 2. As described earlier, this principle predicts that people will alter the doubly-embedded structure by discarding the partial structures directly dependent on the least recent words, until the load associated with the structure is acceptable. Thus, the noun phrase (NP) structure headed by the *man* is discarded. Furthermore, the complementiser phrase (CP) modifying this NP is also discarded, since its existence in this structure is dependent on the existence of the head NP:

[*IP* [*NP* the woman$_j$ [*CP* [*NP* *Oj*] that [*IP*]]]]]]]
X_{TR} X_{TR} X_{LR}

The load associated with this structure is below the maximum allowable load, so the processing can continue without difficulty until the verb *eats* is input.
The verb *eats* cannot attach anywhere in this structure and the parsing fails:

[*IP* [*NP* the woman$_j$ [*CP* [*NP* *Oj*] that [*IP* [*NP* the dog] bit]]] likes]

5.1.3 Using the dual-task paradigm to assess user cost

The experimental questions of this study involve the precise time-course of the expected transition from a facilitatory synchronous multimodal processing to an interfering asynchronous processing. It was decided to use the on-line word-monitoring task to track this transition point in real-time. The task required subjects to press a key on detection of a semantically defined target. As noted in Chapter 2, the logic is that as processing load changes across the sentence, resource allocation to the comprehension task changes as well. Thus, at the point of comprehension difficulty, many resources are devoted to comprehension, so fewer resources are available for the monitoring task. As a result, detection times are slow. The target words appeared either in the second word-position or in the penultimate position in each pair of sentences. Early-position words reflect low-load processing points in both sentence structures. Late-position targets reflect medium-load for the simple structure and high processing-load for the complex structure (always after the breakdown point determined by Gibson's LRNH principle). The use of the two word-positions attempted to reveal preliminary information about the on-line fluctuation of resources, which could be related to the processing of short and long sentences presented to both modalities. An examination of the sentence

comprehension served as a complementary task, to ensure that subjects were processing the sentences for comprehension.

5.2 Experimental hypotheses

5.2.1 Sentence complexity effects on user cost

The MMUM suggests that the excessive storage demands imposed by the need to maintain several unassigned thematic roles in memory will lead to a breakdown during processing of the doubly-embedded sentence structure (c.f., Gibson, 1991). The syntactic parser will fail to maintain the syntactic structure whose processing load is greater than the threshold-value *K*. The model also suggests that the semantic sub-system will face difficulties in interpreting the semantic relationship between constituents, such as the role relationship of the verb, properties of objects and temporal relationships between objects or events. Since these cannot be interpreted on-line, the model assumes that the phonological sub-system will try to maintain word-order information active to enable post-interpretative utilisation of phonological representations. This is bound to fail since referring to the phonological form and the order of lexical items in the sentence can only reinitiate the parsing process and cannot yield meaning directly (Waters et al., 1987).

It was assumed that the word-monitoring technique would prove sensitive enough to reveal on-line processing difficulties of target words that appear after the breakdown point in the doubly-embedded sentences in all presentation conditions. The excessive storage demands imposed by these sentences were expected to slow down the computation of the late-position target words by the lexical, syntactic and semantic systems and thus to increase the average response time in the complex condition. In addition, the failure to maintain the syntactic structure of the doubly-embedded sentences in WM was expected to yield lower comprehension rates than those found for the right-branching sentences in all presentation conditions. This prediction did not aim to validate Gibson's metric but rather, in assuming the validity of the metric, it was expected to demonstrate a systematic manipulation of sentence complexity in this study.

In summary, a main effect of complexity was predicted for both the response time and the comprehension rate measures: lower performance rates for complex than for simple sentences.

5.2.2 Word-position effects on user cost

The experimental design does not enable testing for direct effects of sentence length: the range of sentence lengths is not large enough and the lengths do not distribute in an equal manner between sentences. The design only allows an examination of whether, on average, response times to late-position target words are higher than those found for early-position target words. Nevertheless, sentence length was believed to closely relate to WM limitations, when measuring on-line performance. Following the capacity theory (Just & Carpenter, 1992), it was presumed that each

word requires resources for its activation and integration in the sentence. The total amount of resources used increases as successive words of the sentence are processed. Assuming that resources of sentence processing are limited, then fewer resources are available for the activation of late-position words. This implies that more processing cycles will be needed to reach the recognition threshold of late-position words and, therefore, a longer processing time than that required for early-position words. This hypothesis could only be tested for the auditory-based conditions, where response times could be measured from the onset of the spoken target word. The visual-based conditions only allowed measurement of response times from the onset of the whole sentence with no specific point in time for the onset of processing of the target word. This precluded observation of systematic variation in processing time due to word-position. Still, both modality-based conditions enabled the examination of the interaction between complexity and position. Since early-position targets were placed at low processing-load points in both sentence types, their processing times were not expected to differ across complexity conditions. On the other hand, the higher demands placed upon computation processes at the end of the complex structure were expected to slow lexical-access of late target words in all presentation conditions relative to the simple structure.

In summary, the following effects of position were predicted for the response time measure:

- A main effect of position: slower responses for late than for early target words (only meaningful for the auditory-based conditions).
- An interaction between complexity and position: similar response times for early target words in both complexity conditions, slower responses for late target words in complex than in simple sentences (meaningful testing for both modality-based conditions).

For the comprehension rate measure, higher values were predicted for early-position sentences than for late-position sentences. The language system shares resources with the SAS that (in addition to its multimodal supervision function) coordinates the sentence processing and the word-monitoring tasks in this experiment[21]. An early response should therefore free up resources for the processing of the sentence. Furthermore, both visual-based and auditory-based conditions were expected to show the same effect of position since an early monitoring response should leave more resources for the comprehension of all sentences, regardless of the selected mode of presentation. On the other hand, target position was expected to influence the comprehension of simple and complex sentences in a differential manner: monitoring for late target words was expected to slightly reduce the comprehension of simple sentences, but to severely impair the comprehension of complex sentences. This difference was expected to be larger than the difference between early-simple and early-complex sentences.

Finally, both modality-based conditions (i.e., the two visual-based conditions and the two auditory-based conditions) were expected to show the same pattern of interaction between complexity and position. The rationale for this prediction is as follows: according to the MMUM, the availability of a

[21] The rationale of this contention does not derive from the MMUM, but rather from the dual-task conditions in this experiment.

durable visual text can facilitate slightly the semantic interpretation of long-simple sentences by providing a temporary physical record of word order. In the speech condition, information must be processed in real-time. If processing of the information cannot be completed immediately, then the user must maintain their primary representations for later processing. This implies that under normal processing conditions, the use of regressive eye-movements will diminish the effect of sentence length for simple but not for complex visual-sentences. The word-monitoring task in this study was expected however to reduce regressive eye-movements for late-target sentences. Specifically, for the visual-based conditions, it was believed that the task of monitoring a unique target in each sentence would minimise eye-regressions in late-target sentences, due to resource constraints. For the multimodal auditory-attend (MMAA) condition, it was believed that subjects would prefer not to risk failure in the late-target "catch trials" (see method).

In contrast, for early-target sentences in the visual-based conditions, the execution of the word-monitoring response was expected to leave more resources for both sentence processing and for the control of eye-regressions. Also, an early execution of the monitoring response in the MMAA condition would enable the use of eye-regressions without the risk of failing a trial and, although contradicting the specific instruction to ignore the visual channel, may support the comprehension of simple sentences. Overall, the use of regressive eye-movements in early-target sentences was expected to coincide with the predictions for the position factor[22]. It was expected to support the comprehension of simple sentences in both the visual-based and the auditory-based conditions (the latter include a static-durable visual text presentation in the MMAA condition).

For long-complex sentences, where storage and computation demands exceed the available capacity, the recollection of intermediate representations using eye-regressions was not expected to improve performance. Consistent with Gibson's metric, the MMUM assumes that the syntactic parser will not have sufficient resources available for the delayed assignment of thematic roles required for the comprehension of excessively complex sentences. In spite of this assumption, an early execution of the word-monitoring response was expected to increase the available capacity and therefore to improve comprehension of such sentences. Thus, the source of the predicted facilitation for excessively complex sentences is the early execution of the monitoring response rather than the use of regressive eye-movements.

In summary, the following effects of position were predicted for the comprehension rate measure:

- A main effect of position: higher comprehension rates for early- than for late-target sentences.
- An interaction between complexity and position: lower comprehension rates for late-complex than for early-complex sentences. This difference would be bigger than the difference between late-simple and early-simple sentences.
- No interaction between complexity, position and modality: the interaction between complexity and position would be found in both modality-based conditions.

[22] The effect of regressive eye-movements was not directly examined in this experiment.

- No interaction between modality and position, between multimodality and position or between modality, multimodality and position: executing an early monitoring response should leave more resources for the comprehension of all sentences, regardless of the selected mode of presentation.

5.2.3 Facilitation and interference in multimodal presentation

Visual-based conditions

The visual-based conditions were designed to examine the transition from a facilitatory synchronous processing to an interfering asynchronous processing. In the MMVA condition, subjects were instructed to attend the visual channel and to ignore the auditory channel. However, assuming that speech cannot be ignored (and subjects cannot adhere to this instruction), this presentation condition was intended to provide an approximation of a multimodal presentation in which no manipulation of attention takes place.

As suggested earlier, the MMUM suggests that multimodal facilitation or interference will depend on the extent of synchronisation between the visual and the auditory stimuli. It also proposes that visual and auditory information make their initial contact in the lexical-access system and secondary contact in the phonological store. The time it takes for visual and auditory information to become available to these contact-structures is input dependent. The MMUM does not propose that these structures are always accessed with redundant information. Whereas this may occur for early sentential components, late sentential components will not necessarily make contact on a redundant multimodal basis due to resource constraints. Specifically, although the model assumes that the SAS is capable of supervising the coordination of processing of visual and auditory stimuli by bringing "out of phase" stimuli into phase, it also assumes that long-simple sentences will slightly impair its capability to supervise such coordination. The impairment will be substantially more pronounced for long-complex sentences. The lack of synchronised processing will lead to parallel contact points in the cross-modal sub-systems that are common for both speech and visual text processing in a temporally arbitrary manner.

Based on the above, a triple interaction between complexity, word-position and multimodality was predicted for the response time measure in the visual-based conditions. A reduction in user cost for early targets was expected in the MMVA condition relative to the visual text condition (i.e., redundancy gain). Early targets were placed at a low-load (below threshold) processing point in both simple and complex sentence structures in order to reflect the multimodal activation of the lexical-access system. Late targets were placed at a medium-load processing point in the simple structure and a high-load point (above threshold) in the complex structure. As suggested earlier, the utilisation of resources throughout the complex sentence was assumed to slow down lexical-access of late target words. Since the MMUM assumes that the SAS draws upon the same resources used for both storage and computation for its multimodal supervision process, an increase in user cost for late-complex targets was expected in the MMVA condition relative to the visual text condition (i.e., redundancy cost). A smaller increase was expected for the late-simple targets in the visual-based conditions. In

this case, all systems were expected to compute the sentence successfully, but the storage demands entailed by the sentence length factor were assumed to be sufficient to impair the synchronisation process and to slow down lexical-access of late target words. On average, a significant interaction was predicted between multimodality and position for the visual-based conditions.

In summary, the following effects of multimodality were predicted for the response time measure in the visual-based conditions:

- An interaction between multimodality and position: redundancy gain for early target words, redundancy cost for late target words.
- An interaction between complexity, position and multimodality: a similar redundancy gain for early target words in both complexity conditions, a higher redundancy cost for late target words in complex sentences than in simple sentences.
- No interaction between complexity and multimodality: for simple sentences, similar monitoring times in both visual-based conditions. For complex sentences, a small and insignificant redundancy cost.

For the comprehension rates in the visual-based conditions, a triple interaction was predicted between complexity, position and multimodality. Specifically, whereas for the simple sentences, a similar performance was predicted for both visual-based conditions in each position condition (higher in early- than in late-target sentences), for the complex sentences, a cost increase was predicted for both early- and late-target sentences in the MMVA condition relative to the visual text condition. Also, it was believed that the need to monitor for the appearance of the target in complex sentences would consume resources required by the SAS for its supervision functions of both (i) the competition between the language sub-systems for activation and (ii) the coordination of processing between modalities. Thus, the interference effect for late-target complex sentences was expected to be substantially larger than the predicted effect for early-target complex sentences. This difference was expected to be larger than that predicted for simple sentences. On average, the pattern of this triple interaction was expected to yield a significant interaction between complexity and multimodality for the visual-based conditions: complex sentences presented in the MMVA condition were expected to produce lower comprehension rates than in the visual text condition, whereas simple sentences presented in the MMVA condition were expected to produce a similar performance to the visual text condition.

In summary, the following effects of multimodality were predicted for the comprehension rate measure in the visual-based conditions:

- A triple interaction between complexity, position and multimodality: for simple sentences, similar comprehension rates for both visual-based conditions in each position condition. For complex sentences, an interaction between multimodality and position: a higher redundancy cost for late- than for early-target sentences.
- An interaction between complexity and multimodality: similar comprehension rates in both visual-based conditions for simple sentences, redundancy cost for complex sentences.

- No interaction between multimodality and position: the redundancy cost for late-target sentences would be slightly higher than the expected cost for early-target sentences.

Auditory-based conditions:

As suggested earlier, there is a consensus in the literature that speech cannot be ignored, either at the phonological level, or at the semantic level. A further assumption is sometimes made that, if the speech material is meaningful, any concurrent reading will be impaired (Martin et al., 1988). The MMAA condition allowed the examination of an attended auditory channel in the presence of a static-durable visual text. It was assumed that this control condition would encourage the user to ignore the visual text at points of processing difficulty. Therefore, no indication of multimodal interference was expected in this condition.

Similar to the predictions of the visual-based conditions, a reduction in user cost was predicted for the monitoring times of early target words in the MMAA condition relative to the speech condition. Again, early targets were placed at a low-load (below threshold) processing point in both simple and complex sentence structures and, following Hanson (1981) were expected to reflect the multimodal activation of the lexical-access system. In contrast, for late target words that were placed at a medium-load processing point in simple sentences and a high-load processing point in complex sentences, response times were expected to differ from those of the visual-based conditions. Rather than the predicted redundancy cost, response times for late targets in the MMAA condition were expected to equal those of the speech condition. Moreover, sentence complexity was not expected to differentially affect monitoring times for late target words in the two auditory-based conditions. These predictions were based on the assumption that in contrast to the auditory channel, visual text can be ignored. Therefore, to eliminate the multimodal interference effect at late-position points, the SAS was assumed to activate cognitive operations that lead to focusing attention on the auditory channel.

In summary, the following effects of multimodality were predicted for the response time measure in the auditory-based conditions:

- A main effect of multimodality: faster monitoring times in the MMAA condition than in the speech condition.
- An interaction between multimodality and position: redundancy gain for early target words, similar monitoring times for late target words.
- No interaction between complexity and multimodality: similar redundancy gain in both complexity conditions.
- No interaction between complexity, position and multimodality: a similar redundancy gain for early target words in both complexity conditions, similar response times for late target words in both auditory-based conditions for each complexity condition (higher in the complex than in the simple condition).

The pattern of comprehension rates in the auditory-based conditions was also expected to differ from that of the visual-based conditions. The MMUM proposes that dynamic-transient and static-durable external representations do not involve the same memory demands on users' processing. They differ characteristically in the storage and computational demands determined by different presentation techniques for a successful processing of a given sentence. Significantly, the model suggests that the presence of a static-durable visual text may encourage users to allocate resources to sentence computation rather than to storage of intermediate computational products. This repeated de-allocation of resources may reduce the resources available to keep various intermediate and/or final products of comprehension active in WM during the course of processing a sentence. For complex sentences, where storage and computation demands exceed the available capacity, the unavailability of an earlier piece of information will not allow the parser to compute new elements necessary for the comprehension of the sentence. Consistent with Gibson's metric, the MMUM assumes that regardless of the durability of the presentation condition, the syntactic parser will not have sufficient resources available for the delayed assignment of thematic roles required for the comprehension of complex sentences. Sentence processing therefore stalls and if the user is to continue at all, they must adopt a repair mode possibly requiring the parse to be restarted. For simple sentences, the information collected by regressive eye-movements in the MMAA condition can assist the semantic interpretation of long-simple sentences in that it provides a physical record of word order information. In the speech condition, information must be processed in real-time. However, as suggested earlier, the word-monitoring task in this study was expected to minimise regressive eye-movements for late-target sentences. Similar to the visual-based conditions, it was assumed that regressive eye-movements in the MMAA condition would facilitate the semantic interpretation of early-target simple sentences. Thus, the comprehension of simple sentences in the MMAA condition was expected to be slightly and insignificantly higher than the expected performance in the speech condition. For complex sentences, a similar performance was predicted for both auditory-based conditions (a higher performance than that predicted for the MMVA condition).

The suggestion that the MMAA condition may improve performance of complex sentences, relative to the auditory only condition, was rejected. Theoretically, the durable visual text in the MMAA condition enables the user to use eye regressions to recollect early sentential components while attending to the spoken continuation of the sentence. The model predicts that this is not expected to alleviate the high processing cost of long-complex sentences. This prediction is based on the assumption that the SAS will not have sufficient resources available for its supervision of (i) the competition between the language sub-systems for activation and (ii) the assignment of thematic roles across modalities.

In summary, the following effects of multimodality were predicted for the comprehension rate measure in the auditory-based conditions:

- No interaction between multimodality and position: similar comprehension rates in both auditory-based conditions for each position condition (higher for early- than for late-target sentences).

- No interaction between complexity and multimodality: similar comprehension rates in both auditory-based conditions for each complexity condition (higher for simple than for complex sentences).
- No interaction between complexity, position and multimodality: similar comprehension rates in both auditory-based conditions for each position condition in both complexity conditions.

Overall, a significant interaction between complexity, modality and multimodality was predicted for the comprehension rate measure: for simple sentences, the attended channel will not affect comprehension significantly in either modality-based condition. For complex sentences, the unattended speech will impair reading comprehension in the visual-based conditions, whereas the unattended visual text will not affect speech processing in the auditory-based conditions.

Furthermore, the expected speech interference in comprehending complex sentences was expected to be large enough to produce a significant redundancy cost when calculated across modalities. This was expected to produce a significant interaction between complexity and multimodality: no effect of multimodality in comprehending simple sentences and a multimodal interference in comprehending complex sentences.

Complexity was also expected to differentially affect comprehension rates in the two modality-based conditions. Comprehension of simple sentences will be slightly higher in the visual-based conditions than in the auditory-based conditions. On the other hand, assuming that added speech impairs comprehension of complex sentences in the MMVA condition and that added text can be ignored at points of processing difficulty in the MMAA condition, comprehension of complex sentences will be substantially lower in the visual-based conditions than the auditory-based conditions.

On average, modality was expected to interact with multimodality. The added speech was expected to impair comprehension in the visual-based conditions, whereas the added visual text was not expected to affect comprehension in the auditory-based conditions. Finally, neither the main effect of modality, nor the main effect of multimodality was expected to reach significance in this study.

5.2.4 Individual differences

The MMUM hypothesises that individual differences in verbal WM capacity will affect performance when the combination of the multimodal presentation technique and the linguistic complexity of the presented materials imposes a high processing load on the users. The higher the processing load, the more apparent will be the difference. Predictions for the effect of WM capacity on reaction time and comprehension measures were discarded due to the unreliability of the measure of WM capacity used in this study.

5.3 Method

5.3.1 Materials and design

Appendix A presents the materials used in this experiment. The primary stimulus set consisted of 88 pairs of sentences, each containing one right-branching sentence and one doubly-embedded sentence. Each pair of sentences contained a unique monitoring target word (see underlined words in Appendix A). The 88 target words were all adjectives selected by three judges, as their concreteness and imaginability values could not always be identified by standard methods (e.g., the Pavio norms (Pavio, Yuille & Madigan, 1968), the Colerado norms (Toglia & Battig, 1978), and the Gilhooly-Logie norms (Gilhooly & Logie, 1980)). Each word was required to be a frequent response to the category in question. Words from 16 taxonomic categories were selected (e.g., category: colour; target: *red*). Target words were assigned either in the second word-position (early position) or in the penultimate position (late position) in each pair of sentences.

Subjects were assigned to one of the two experimental groups created by the complexity factor[23]. Each complexity group was presented with the sentences in four presentation conditions[24]:
1. (Static-durable) visual text
2. Speech

---

[23] With so many factors examined, a decision had to be made about which variable would form the between subjects variable in this experiment. The choice was guided by which independent variables were considered more sensitive, stable or fundamental. The complexity variable was considered ideal despite the implication that participants would experience only one syntactic style, which in turn - might increase their use of strategies. The reasons are numerous: first, both modality and multimodality were expected to be quite sensitive to the effect of between subjects variance and were considered better as within subject variables. In contrast, the complexity variable was considered more stable and therefore more suitable to be a between subjects variable. The second, more important, reason is that using complexity as a within subjects variable would have meant comparing between simple and complex sentences that differ in their meaning. Choosing syntactic complexity as the between subjects variable enabled the experiment to assess the effects of syntactic complexity on the user's ability to coordinate between visual and auditory information, using sentence pairs of the same meaning. In addition, adding further syntactic styles was not a viable option. The introduction of further styles for the given 88 sentences was rejected on the ground of compromising the experimental power, whereas adding more sentences was considered inappropriate; the subjects had to process 88 sentences and it was not possible to prolong the experiment by adding further sentences without making them tired. It was believed that although strategies were still likely, the experimental design would minimise them. For example, the use of minimal intonation and prosody intended to equate between the visual and the auditory conditions but also to minimise the use of recall strategies in the simple condition. The use of excessively complex sentences was believed to eliminate finding out the canonical order of the agents and the themes in the sentences in spite of the sole use of the doubly-embedded structure.

[24] Presentation was varied across uniform blocks, since in the multimodal conditions subjects were asked to attend one modality and to ignore the other (See section 5.1.1). Having presentation forms mixed would have rendered this instruction meaningless.

3. (Static-durable) multimodal visual-attend (MMVA)
4. (Static-durable) multimodal auditory-attend (MMAA)

Digital recording and editing of speech was conducted using the Sample Editor application (sample rate 22.255 kHz, sample size 8 bits). Sentences were recorded and edited so that (i) the target word of both the simple and the complex versions in a single pair of sentences had the same onset-duration and (ii) both sentence versions had the same presentation duration. The sentences were spoken in a male voice, with minimum intonation and prosody to equate between the visual and the auditory conditions (the visual conditions did not include commas). They varied in length (from 12 to 15 words) and had an average presentation rate of 204 words per minute (the visual display of text could not be terminated early by the subject in order to advance more quickly to the true/false question). Although slightly higher than average speech rate, and although relatively high for the right-branching and the doubly-embedded structures used, this presentation rate conformed with the normal reading speed as defined by Masson (1985), (see Chapter 2, section 2.3.1). Significantly, it was assumed that a slower presentation rate would allow subjects to visually scan for late target words, especially in the simple conditions. Presentation duration of each sentence took on average 4074 ms.

A visual text version was created for 66 sentences using Palatino 18 points font. Visual sentence pairs were selected and allocated to the three presentation conditions for each complexity group that included visual output: visual text, MMVA, and MMAA. 66 recorded sentence pairs were selected and allocated to the three presentation conditions for each complexity group that included spoken output: speech, MMAA, and MMVA.

Catch trials

Four sentences in each presentation condition constituted the catch trials. Catch trial sentences included an adjective that did not match the word category provided for each sentence. In the single modal conditions (speech, visual text), this enabled an examination of whether subjects were responding to the actual target words or were simply guessing. In the catch trials of the multimodal presentation conditions, the visual sentences differed from the spoken sentences only with respect to the target word. The target words appeared only in the channel which should be ignored, whereas the channel which should be attended to included a different adjective that did not match the pre-specified word category, but that was appropriate to the context of the sentence. This served a different purpose: it was intended to make sure that subjects would attend to the required channel (i.e., the visual channel in the MMVA condition and auditory channel in the MMAA condition). In the absence of this control, one could not interpret the experimental results in a meaningful way (to determine the source of facilitation or interference for each of the multimodal conditions).

5.3.2 Comprehension statements

A "true" or "false" comprehension statement was created for each sentence (including catch trials), to assess whether subjects were properly processing each presented sentence. The comprehension test items for each sentence were constructed by combining one of the three verbs with two of the four nouns (e.g., *The man that the woman that the dog bit likes eats red meat*; statement: *The dog bit the man* – false). Pragmatically meaningless items were excluded.

5.3.3 Implementation

The implementation of these materials was made using the SuperLab application. For each complexity group, presentation conditions were grouped in four different blocks. Each block consisted of 22 sentences, half of which included an early target word and the other half a late target word. Two of the early and two of the late-target sentences of each presentation condition were catch trials.

Sentences were balanced across presentation conditions with respect to their pragmatic complexity (an index of content)[25], their length (number of words) and the serial position of the target words relative to the sentence length. In addition, the comprehension statements were balanced across presentation conditions with respect to the number of true/false statements[26] and their internal distribution of the nouns-verb combinations. The presentation of early and late-position target sentences was randomised within each presentation block.

Four experimental versions were created for each complexity group; each having a different order of the presentation-conditions blocks to minimise order effects[27]:

---

[25] This step intended to make the sentences in the four presentation conditions equally easy to guess (see Chapter 2, section 2.2.4). The sentences themselves were not randomly allocated to the different presentation conditions to avoid a situation in which one condition has, on average, higher pragmatic complexity values than another. It is correct that counterbalancing sentences across presentation conditions could have been controlled for the pragmatic complexity aspect (i.e., using systematic allocation of sentences that are equal in their pragmatic complexity values across conditions), but this implied having too many experimental versions. In each experiment, four order versions were created for each complexity group and each served four subjects. Having more versions would have meant having fewer subjects in each condition. This seemed quite extreme for an experiment with 32 subjects. Since the contents used are from mundane domains and therefore assumed equivalent, balancing sentences across presentation conditions with respect to their pragmatic complexity value was considered a sufficient measure of control.

[26] The truth-value of the statements was balanced across presentation conditions in order not to bias subjects' responses towards any specific value. Yet, the actual truth-value of the statements was not assumed to affect performance rates. This is why the truth-value of the statements was not extracted in the statistical analysis. Even if performance differences exist when comprehension statements are true than when they are false, balancing this factor means that they distribute equally across presentation conditions.

[27] According to Howell (1997), systematic ordering of conditions distributes order effects across the cells of the design, lumping them into the error term.

Version 1: visual text, MMVA, speech, and MMAA
Version 2: speech, MMAA, visual text and MMVA
Version 3: MMVA, visual text, MMAA and speech
Version 4: MMAA, speech, MMVA and visual text

A further set of sentences was constructed, to serve as practice materials for the subjects. These contained examples of all presentation conditions including catch trials.

5.3.4 Apparatus

The experiment was run on a PowerPC 7200/90. Spoken sentences were presented at a comfortable volume through a pair of speakers, and the visual material was presented on a standard Apple display, size 15 inch. Subjects' responses were collected via Apple Desktop Bus™ (ADB) keyboard that can be accessed quickly (in about 4 ms) by the Time Manager 1 of the SuperLab application. This provides a higher resolution (4 ms or better) than the one provided by the normal Macintosh tick counts (± 16.67 ms).

5.3.5 Procedure

Subjects first read a comprehensive set of instructions, describing the experimental tasks and the sentence structure they would encounter, illustrated with an example. They were told that their task consisted of two parts: (i) to identify a particular word in the sentence and (ii) to judge a true or false statement, which would test their understanding of the sentence. Subjects were also told the order in which they would have to perform these tasks: that before each sentence was presented, they would be given a category (for example: "Colour"), for which they would have to monitor for an adjective belonging to that category. They were also told that they should respond to the adjective (by pressing the "0" key on the lower right side of the keyboard) as soon as they encountered it while processing the sentence. The instructions included a description of the catch trials, to which subjects were requested to respond using a different key ("2"). They also included a warning that pressing the "0" key during a catch trial or the "2" key during a regular trial would produce an alert sound (a "Quack"). Subjects were also told that after the sentence had disappeared, they would be presented with a statement, which they would have to judge as true or false (by pressing the "Z" key for a true statement and the "X" key for a false statement); an alert sound (a "Beep") would be given for an erroneous response. Finally, subjects were instructed to respond as quickly and accurately as they could. The need for speedy responses to the target words was also emphasised verbally by the experimenter.

After confirming that the instructions were understood, the practice session started. The practice was divided into the 4 different presentation conditions described above. Subjects were presented with specific instructions for each of the presentation conditions followed by the practice sentences.

Subjects that showed difficulties during practice had to repeat the session until a satisfactory level of performance was reached prior to the start of the experiment.

Experiment 1 concluded with the reading span test, described next.

5.3.6 Span task

The reading span test is originally based on reading aloud sentences from separate index cards. The task was computerised to make its administration more efficient. Subjects read sets of sentences aloud, one sentence at a time and then recalled each of the sentence final words from that set. Subjects were given sets of an increasing number of sentences until they failed to recall all final words for three out of the five sets at a particular level. Reading span was defined according to this stopping rule. If a subject showed perfect recall in two of five groups of a given set size, then half credit for that size was given. The test sentences in each group were 13 to 16 words long and were unrelated to each other (see Daneman & Carpenter, 1980, 1983, for a further description of the test).

According to the original method, subjects whose reading spans are 3 or 3.5 should be classified as *Medium* span subjects and excluded from any span-based analysis. Unfortunately, the available resources necessitated using their data in all experiments. All subjects were classified according to their performance as being either *High* span or *Low* span. High span subjects were those whose reading spans were 3.5 or higher, while low span subjects were those whose reading spans were 3.0 or lower.

5.3.7 Subjects

The subjects were 30 students and faculty members at City University, who were paid £6 for their participation. English was required to be the first language of all subjects. All reported normal or corrected to normal acuity and normal hearing. 7 subjects were excluded from the analysis: 2 were run prior to a change in the instructions, and 5 showed a high proportion of missed catch trials or falsely recognised catch trials, and were excluded from the analysis. Subjects were randomly assigned to one of the versions of the two experimental groups created by the syntactic complexity factor and the different ordering of the presentation condition.

5.4 Results

Unfortunately, administering the span test after the experimental session proved to be a methodological error. Subjects in the complex condition were too tired to reach a high performance in the span test even if their performance in the experimental session was good. A preliminary exploration revealed that for the complex sentences, low span subjects reached higher comprehension rates than high span subjects. The span measure was therefore deemed to be unreliable and was eliminated from all analyses.

5.4.1 Effects of complexity

Separate analyses of variance were conducted for the word-monitoring data of the two visual-based and auditory-based conditions. The overall means for the two complexity conditions by the two presentation conditions and the two word-positions in the visual-based conditions are given in Table 5.2. The analysis of the visual-based conditions indicates that the main effect of complexity failed to reach significance ($F(1, 21) = .001$). In addition, complexity failed to interact with position ($F(1, 21) = .832$) and with both position and multimodality ($F(1, 21) = .296$).

Table 5.2

Mean Response Time (RT) for the Visual-Based Conditions (msec): By Complexity, Multimodality and Word-Position (Standard Errors in Parentheses)

	Presentation Conditions by Position			
Complexity Level	Visual-Early	Visual-Late	MMVA-Early	MMVA-Late
Simple	1801	3018	1514	2922
	(201)	(239)	(205)	(183)
Complex	1626	3018	1438	3148
	(193)	(229)	(197)	(176)

Similar results were obtained in the two analyses conducted for the word-monitoring task in the auditory-based conditions. The first analysis was conducted for the absolute response times of the word-monitoring task (RT - measured from the onset of the sentence). The overall means for the two complexity conditions by multimodality and position in the auditory-based conditions are given in Table 5.3. The second analysis was conducted for the relative response times of the monitoring task (RT' - measured from the onset of the spoken target). The data is presented in Table 5.4. Neither of the analyses yielded significant results: in both analyses, the main effect of complexity failed to reach significance (RT: $F(1, 21) = .022$). In addition, complexity failed to interact with position (RT: $F(1, 21) = .275$; RT': $F(1, 21) = .450$).

Table 5.3

Mean Absolute Response Time (RT) for the Auditory-Based Conditions (msec): By Complexity, Multimodality and Word-Position (Standard Errors in Parentheses)

	Presentation Conditions by Position			
Complexity Level	Speech-Early	Speech-Late	MMAA-Early	MMAA-Late
Simple	1254	4148	1683	3825
	(104)	(79)	(242)	(71)
Complex	1217	4202	1642	3892
	(99)	(75)	(232)	(68)

Table 5.4

Mean Relative Response Time (RT') for the Auditory-Based Conditions (msec): By Complexity, Multimodality and Word-Position (Standard Errors in Parentheses)

	Presentation Conditions by Position			
Complexity Level	Speech-Early	Speech-Late	MMAA-Early	MMAA-Late
Simple	1158	961	1590	715
	(103)	(78)	(242)	(72)
Complex	1120	1042	1547	804
	(99)	(75)	(232)	(69)

These results suggest that the word-monitoring task did not prove sensitive enough to (i) demonstrate the on-line processing difficulties of late target words in complex sentences or to (ii) demonstrate that the excessive storage demands of complex sentences exhaust the limited-pool of resources also used by the SAS to supervise the coordination between the visual and the auditory information channels.

For the comprehension rate measure, a combined analysis of variance was conducted for the two modality-based conditions[28],[29]. Table 5.5 provides the mean comprehension rates found for the visual-based conditions by complexity, position and multimodality. Table 5.6 provides the means obtained in the auditory-based conditions.

[28] An exploration of the comprehension rates data revealed that the sub-conditions created by the complexity, modality, multimodality and position variables did not exhibit perfectly normal distributions. The assumption of normality was not met for complex sentences in the MMAA-Early condition (Shapiro-Wilk (12) = .852; p = .043) and for the simple sentences in the MMAA-Late condition (Shapiro-Wilk (11) = .834; p = .036). Furthermore, their distributions varied in shape: 7 sub-conditions were positively skewed and 9 were negatively skewed. When squared values were used, all distributions were skewed in the same direction. This transformation however, further impaired the normality values of the above conditions and the normality value of the Speech-Early simple condition (Shapiro-Wilk (11) = .820; p = .023). It was decided to report the original values of the comprehension rate measure rather than the transformed values, although the latter produced more significant results in the analysis of variance.

[29] An additional analysis was conducted at the request of the examiners of this dissertation. This analysis included the order in which subjects performed the four presentation conditions as an additional between-subjects variable (called version number) to make sure that the experimental results were unaffected by order effects (see Method, section 5.3.3). None of the results reported in this section was affected by the order in which subjects performed the four presentation conditions. Note that the number of subjects in each version is too small to make this a reliable conclusion.

Table 5.5

Mean Comprehension Rate (CR) for the Visual-Based Conditions (%): By Multimodality, Word-Position and Complexity (Standard Errors in Parentheses)

	Presentation Conditions by Position			
Complexity Level	Visual-Early	Visual-Late	MMVA-Early	MMVA-Late
Simple	71%	78%	71%	73%
	(5%)	(5%)	(6%)	(4%)
Complex	57%	58%	67%	64%
	(5%)	(5%)	(5%)	(4%)

Table 5.6

Mean Comprehension Rate (CR) for the Auditory-Based Conditions (%): By Multimodality, Word-Position and Complexity (Standard Errors in Parentheses)

	Presentation Conditions by Position			
Complexity Level	Speech-Early	Speech-Late	MMAA-Early	MMAA-Late
Simple	65%	59%	81%	72%
	(6%)	(3%)	(5%)	(5%)
Complex	55%	51%	65%	58%
	(6%)	(3%)	(4%)	(5%)

This measure yielded a different pattern of results: overall, the average comprehension of simple sentences (71%) was 12% higher than that found for the complex sentences (59%), yielding a main effect of complexity ($F(1, 21) = 8.259$; $p < .01$). The manipulation of sentence complexity was therefore successful.

5.4.2 Effects of word-positions

Since each visual sentence was fully available from the onset of sentence presentation, response times collected for the word-monitoring task in the visual-based conditions could not be measured from the onset of the visual targets. The absolute response times reveal that late-position targets (3026 ms) were responded to more slowly than early-position targets (1595 ms); ($F(1, 21) = 119.280$; $p < .01$). The same result was found for the auditory-based conditions: the absolute response times reveal that late-position targets (4017 ms) received slower responses than early-position targets (1449 ms); (RT: $F(1, 21) = 725.069$; $p < .01$); (see Table 5.7). These results simply confirm that subjects were processing visual text in a serial and orthodox fashion. Contrary to the prediction, response times measured from the onset of the spoken targets were significantly higher for early-position targets (1353 ms) than for late-position targets (882 ms); (RT': $F(1, 21) = 25.706$; $p < .01$).

Table 5.7

Mean Absolute Response Time (RT) for the two Modality-Based Conditions (msec): By Word-Position (Standard Errors in Parentheses)

	Modality-based Conditions	
Word-Position	Visual-based	Auditory-based
Early	1595	1449
	(123)	(115)
Late	3026	4017
	(133)	(47)

The comprehension rates data, examined for effects of position, does not support the predictions of the MMUM. Overall, comprehension rates of early-position sentences (66%) did not differ from those found for the late-position sentences (64%) ($F(1, 21) = 1.270$). Moreover, complexity failed to interact with position ($F(1, 21) = .110$). The absence of a significant triple interaction between complexity, position and modality ($F(1, 21) = 1.027$) indicates that the two modality-based conditions did not differ in this respect. The overall comprehension rates for the two modality-based conditions by complexity and position are given in Table 5.8. The relationships between these means can be seen more readily in Figure 5.1.

Table 5.8

Mean Comprehension Rate (CR) for the two Modality-Based Conditions (%): By Complexity and Word-Position (Standard Errors in Parentheses)

	Modality-based Conditions by Position			
Complexity Level	Visual-based Early	Visual-based Late	Auditory-based Early	Auditory-based Late
Simple	71%	75%	73%	65%
	(4%)	(4%)	(4%)	(4%)
Complex	62%	61%	60%	55%
	(4%)	(4%)	(4%)	(4%)

Figure 5.1

Mean Comprehension Rate (%): By Modality-Based Conditions, Word-Position and Complexity Conditions

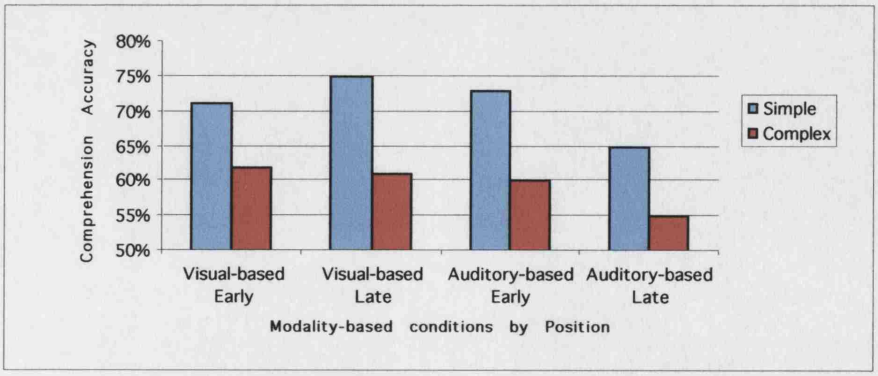


Figure 5.1 shows that the position of the target words affected the comprehension rates differentially in the two modality-based conditions. The significant interaction between modality and position ($F(1, 21) = 5.137$; $p < .04$) is clearly demonstrated by Table 5.9 and Figure 5.2.

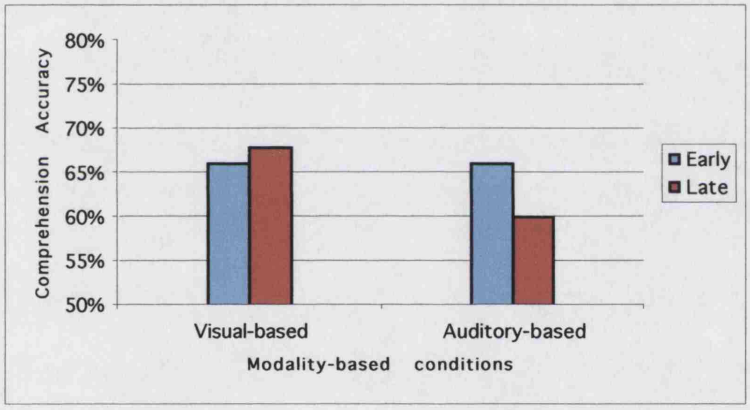
Table 5.9

Mean Comprehension Rate (CR) for the two Modality-Based Conditions (%): By Word-Position (Standard Errors in Parentheses)

	Modality-based Conditions		
Word-Position	Visual-based	Auditory-based	Mean difference
Early	66%	66%	0%
	(3%)	(3%)	
Late	68%	60%	8%
	(3%)	(3%)	
Mean difference	-2%	6%	

Figure 5.2

Mean Comprehension Rate (%): By Word-Position and Modality-Based Conditions



As expected, responses made in the auditory-based conditions for the early-target sentences, yielded higher comprehension rates (66%) than those found for the late-target sentences (60%) ($F(1, 21) = 7.509$; $p < .02$). In contrast, comprehension rates in the visual-based conditions were not affected by the timing of the execution of response (66% for early-position sentences and 68% for late-position sentences), as indicated by the absence of a main effect of position ($F(1, 21) = .477$). Finally, the absence of a significant interaction between modality, multimodality and position ($F(1, 21) = .081$) suggests that the unimodal and the multimodal conditions did not differ in this respect in each modality-based condition (see Table 5.10 and Figure 5.3).

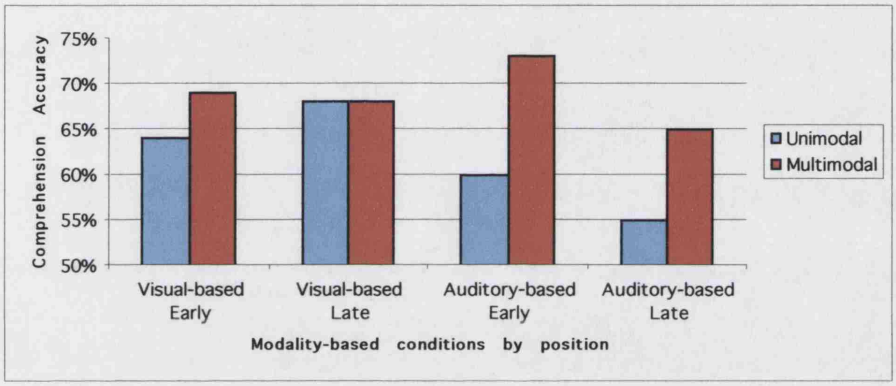
Table 5.10

Mean Comprehension Rate (CR) for the two Modality-Based Conditions (%): By Multimodality and Word-Position (Standard Errors in Parentheses)

Multimodality	Modality-Based Conditions by Position			
	Visual-based Early	Visual-based Late	Auditory-based Early	Auditory-based Late
Unimodal	64% (3%)	68% (3%)	60% (4%)	55% (2%)
Multimodal	69% (4%)	68% (3%)	73% (3%)	65% (5%)

Figure 5.3

Mean Comprehension Rate (CR) for the two Modality-Based Conditions (%): By Multimodality and Word-Position



5.4.3 Effects of Multimodality

The results of the separate analyses conducted for the mean response times of the two modality-based conditions were described earlier. In either analysis, complexity failed to interact with the position variable, or with position and multimodality. It was suggested that the word-monitoring technique did not prove sufficiently sensitive to test the assumptions about on-line processing in relation to resource demands and the effects of presentation. On the other hand, each of the analyses revealed a significant interaction between the multimodality and the position variables. Table 5.11 provides the mean response times found for the two presentation conditions by the two word-positions in the visual-based conditions. Figure 5.4 presents the relationships between these means.

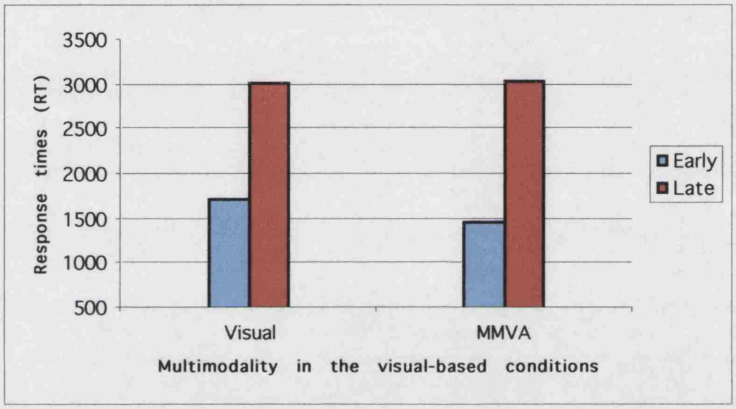
Table 5.11

Mean Response Time (RT) for the Visual-Based Conditions (msec): By Multimodality and Word-Position (Standard Errors in Parentheses)

	Multimodality Conditions		
Word-Position	Visual	MMVA	Mean difference
Early	1714	1476	238
	(139)	(142)	
Late	3018	3035	-17
	(165)	(127)	
Mean difference	1304	1559	

Figure 5.4

Mean Response Time (RT) for the Visual-Based Conditions (msec): By Multimodality and Word-Position



The significant interaction ($F(1, 21) = 4.710$; $p < .05$), demonstrates the expected redundancy gain for early-position targets in the MMVA condition. This gain of 238 ms is approaching significance ($F(1, 21) = 3.158$; $p = .089$). However, Figure 5.4 shows that response times to late-position targets did not differ across presentation conditions ($F(1, 21) = .031$). On average, the results indicate slightly higher response times to target words in the visual condition (2366 ms) than in the MMVA condition (2255ms). This small difference of 111 ms failed to yield a significant main effect of multimodality ($F(1, 21) = .901$).

Table 5.12 and Figure 5.5 provide the mean absolute response times found for the two presentation conditions by the two word-positions in the auditory-based conditions. The analysis conducted for these absolute response times yielded a significant interaction between multimodality and position (RT: $F(1, 21) = 35.973$; $p < .01$). This interaction demonstrates a surprising cost of 427 ms for early-position targets in the MMAA condition (RT: $F(1, 21) = 14.554$; $p < .01$). It also demonstrates a redundancy gain of 316 ms for late-position targets (RT: $F(1, 21) = 55.458$; $p < .01$). On average, the results indicate slightly higher response times to target words in the MMAA condition (2760 ms) than in the auditory condition (2705 ms). These 55 ms failed to yield a main effect of multimodality (RT: $F(1, 21) = .836$).

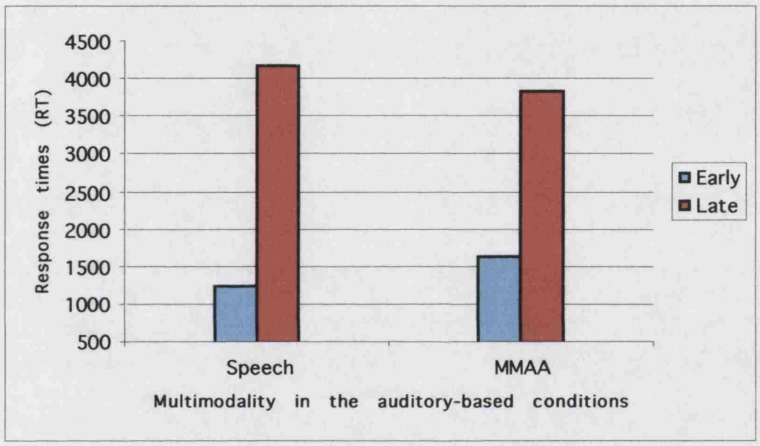
Table 5.12

Mean Absolute Response Time (RT) for the Auditory-Based Conditions (msec): By Multimodality and Word-Position (Standard Errors in Parentheses)

	Multimodality Conditions		
Word-Position	Speech	MMAA	Mean difference
Early	1235	1662	-427
	(72)	(167)	
Late	4175	3859	316
	(54)	(49)	
Mean difference	2940	2197	

Figure 5.5

Mean Absolute Response Time (RT) for the Auditory-Based Conditions (msec): By Multimodality and Word-Position



Similar results were obtained when response times were measured from the onset of the spoken targets (they are shown in Table 5.13 and Figure 5.6). The significant interaction between multimodality and position (RT': $F(1, 21) = 31.716$; $p < .01$) clearly demonstrates a redundancy cost of 430 ms for early-position targets in the MMAA condition (RT': $F(1, 21) = 14.631$; $p < .01$). Similarly, the results demonstrate a redundancy gain of 243 ms for late-position targets (RT': $F(1, 21) = 42.863$; $p < .01$). Again, the overall results indicate slightly higher response times to target words in the MMAA condition (1165 ms) than in the speech condition (1071 ms). These 94 ms failed to yield a main effect of multimodality (RT': $F(1, 21) = 2.324$).

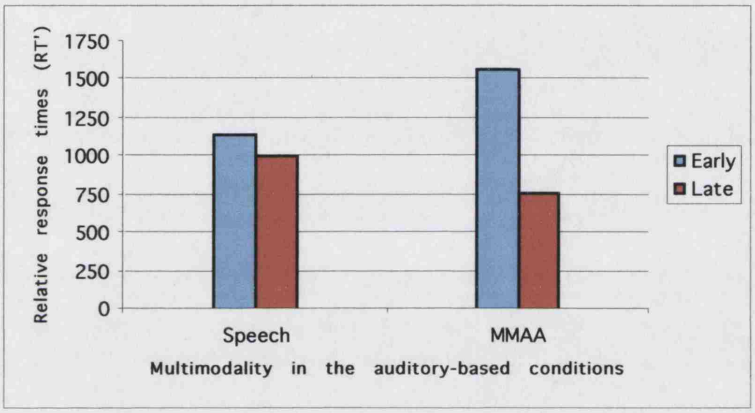
Table 5.13

Mean Relative Response Time (RT') for the Auditory-Based Conditions (msec): By Multimodality and Word-Position (Standard Errors in Parentheses)

	Multimodality Conditions		
Word-Position	Speech	MMAA	Mean difference
Early	1138	1568	-430
	(72)	(168)	
Late	1004	761	243
	(54)	(50)	
Mean difference	134	807	

Figure 5.6

Mean Relative Response Time (RT') for the Auditory-Based Conditions (msec): By Multimodality and Word-Position



The combined analysis of variance, conducted for the comprehension rates of the two modality-based conditions provides the most interesting result of this experiment: an approaching significance interaction between complexity, modality and multimodality ($F(1, 21) = 3.955$; $p = .060$). Table 5.14 provides the mean comprehension rates of all presentation conditions by complexity and multimodality. Figure 5.7 demonstrates the complex relationships between these variables.

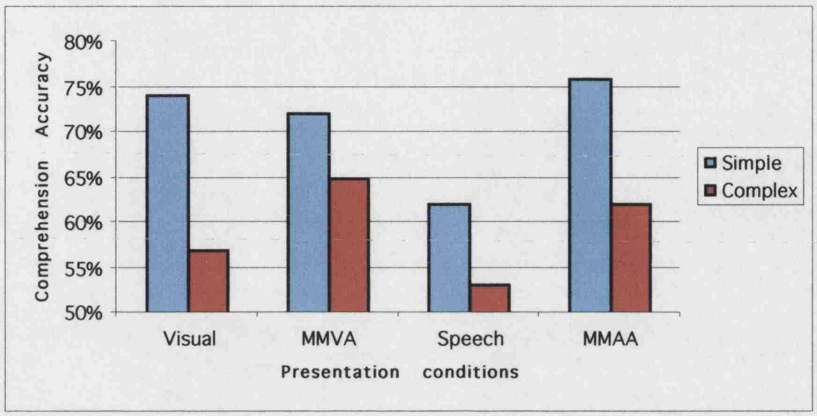
Table 5.14

Mean Comprehension Rate (CR) for the two Modality-Based Conditions (%): By Complexity and Multimodality (Standard Errors in Parentheses)

	Presentation Conditions				
Complexity Level	Visual	MMVA	Speech	MMAA	Mean
Simple	74%	72%	62%	76%	71%
	(4%)	(4%)	(4%)	(5%)	
Complex	57%	65%	53%	62%	59%
	(4%)	(4%)	(4%)	(4%)	
Mean	66%	69%	57%	69%	65%

Figure 5.7

Mean Comprehension Rate (CR) for the two Modality-Based Conditions (%): By Complexity and Multimodality



The investigation of this triple interaction involved a reduced analysis of variance at each level of the modality variable. For the visual-based conditions, the main effect of complexity reached significance ($F(1, 21) = 5.464$; $p < .04$). Of greater interest, the interaction between complexity and multimodality was significant ($F(1, 21) = 5.169$; $p < .04$). Figure 5.7 shows that, as expected, multimodality does not improve comprehension of simple sentences ($F(1, 21) = .587$). On the other hand, rather than the expected redundancy cost, a redundancy *gain* of 8% was found for the complex sentences. The effect reached significance, indicating that a static-durable multimodal presentation (65%) is superior to a static-durable visual presentation (57%) of such sentences ($F(1, 21) = 6.175$; $p < .04$). Moreover, a One-Sample T-Test revealed that comprehension rates of complex sentences in the visual text condition did not depart from a chance level of performance (i.e., from the value of .50), ($t_{11} = 1.773$, $p > .05$) whereas those in the MMVA condition were higher than chance ($t_{11} = 3.527$, $p < .05$). The addition of speech to text made the comprehension of excessively complex sentences possible. Overall, the pattern of results did not yield a main effect of multimodality in the visual-based conditions ($F(1, 21) = 1.366$).

For the auditory-based conditions, the main effect of complexity reached significance as predicted ($F(1, 21) = 5.478$; $p < .04$). Furthermore, consistent with the experimental predictions, complexity and multimodality did not interact ($F(1, 21) = 1.244$). However, the pattern of results is different than expected: rather than the similar comprehension rates predicted for both auditory-based conditions in each complexity condition, it appears that multimodality improves comprehension for both levels of syntactic complexity ($F(1, 21) = 19.970$; $p < .01$). For the complex sentences, it also makes them to depart from a chance level of performance. A One-Sample T-Test revealed that similar to the visual-based conditions, comprehension rates in the speech condition (53%) did not differ from chance ($t_{11} = .789$, $p > .05$) whereas those in the MMVA condition (62%) were higher than a chance level of performance ($t_{11} = 2.803$, $p < .05$).

The absence of a main effect of multimodality in the visual-based conditions and the presence of the effect in the auditory-based conditions yielded a significant interaction between modality and multimodality in the analysis of the comprehension rates ($F(1, 21) = 4.908$ $p < .05$); (see Table 5.15 and Figure 5.8).

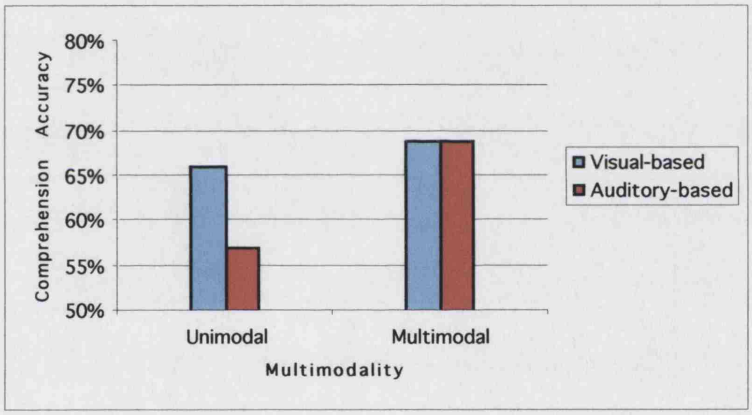
Table 5.15

Mean Comprehension Rate (CR) for all Complexity Conditions (%): By Modality and Multimodality (Standard Errors in Parentheses)

	Multimodality Conditions		
Modality-based Conditions	Unimodal	Multimodal	Mean difference
Visual-based	66%	69%	3%
	(3%)	(3%)	
Auditory-based	57%	69%	12%
	(3%)	(3%)	
Mean difference	9%	0%	

Figure 5.8

Mean Comprehension Rate (CR) for all Complexity Conditions (%): By Modality and Multimodality



Finally, the pattern of results failed to yield the expected interaction between complexity and multimodality ($F(1, 21) = .686$). Overall, the average comprehension rates in the multimodal conditions (69%) were 7% higher than those found in the unimodal conditions (62%), yielding a significant main effect of multimodality ($F(1, 21) = 27.510$; $p < .01$). The main effect of modality (4% advantage found for the visual-based conditions over the auditory-based conditions) was not significant ($F(1, 21) = 1.994$).

5.4.4 Catch trials

Catch trial sentences were the means by which manipulation of attention was controlled in this study. Specifically, they aimed to examine whether subjects responded to the correct target words in the unimodal conditions (rather than arbitrarily) and also whether subjects were attending to the correct

modality and ignoring the other modality as requested. These two different purposes of control reflect the different nature of the catch trials in the unimodal and multimodal conditions. In the unimodal conditions, catch trial sentences did not include an appropriate target word. The position variable was meaningless in these conditions, since subjects had to process a complete catch trial sentence in order to decide that it did not contain a compatible adjective. Missing a catch trial in the unimodal conditions simply meant responding to the wrong target. On the other hand, in the multimodal conditions the target always appeared in the unattended channel and the distractor always appeared in the attended channel. Missing catch-trials indicated an allocation of attention to the wrong channel. The analysis of responses for the catch trial sentences is presented in Appendix B and its main findings are outlined next.

Results indicate that subjects missed more present catch trials in the multimodal than in the unimodal conditions. Since speech cannot be ignored, this finding is not surprising for the visual-based conditions: subjects faced difficulties in ignoring the targets when they appeared in the "unattended" auditory channel that, in practice, cannot be ignored. However, the absence of a main effect of modality in the multimodal conditions indicates that the visual text was not ignored in the MMAA conditions either. Therefore, manipulation of attention in this study was at least partially unsuccessful.

Another analysis was conducted for the recognition rate of target words in the regular trials. The motivation of this analysis was to find out whether the word-monitoring task was equal in ease in all conditions. Any differences in ease of response would reflect on the overall pattern of results reported throughout this section. Results suggest that the recognition rate of target words in the regular trials was significantly higher in the multimodal conditions relative to the unimodal conditions. This implies that regardless of sentence complexity, it was more difficult to correctly capture target words for regular trials in the unimodal conditions relative to the multimodal conditions in both modality-based conditions. Moreover, the view of difficulties in the unimodal conditions gained support from an inspection of subjects' data. This inspection revealed that in each of the unimodal conditions, there were two target words which approximately third of all subjects mistook for catch trials ('famous' and 'light' in the visual condition, 'wide' and 'little' in the speech condition). The rate of mistakes did not appear to be affected by the complexity of the presented sentences. No indication of specific difficulties was found for the multimodal conditions. It is therefore proposed that the word-monitoring task was more difficult in the unimodal conditions relative to the multimodal conditions. This should be kept in mind while analysing the global pattern of results of this experiment.

5.5 Discussion

The information that can be inferred from the word-monitoring data and the possible strategies that could account for the complex pattern of the comprehension data are discussed next. The discussion includes the effects of syntactic complexity, word-position, the attended modality and multimodality on user cost.

5.5.1 Sentence complexity

The results indicate that the experimental manipulation failed to capture the on-line processing dynamics, particularly as they relate to syntactic complexity. The fundamental weakness was the use of the word-monitoring task in conjunction with the task of reading visual text for meaning, using a static-durable presentation of the visual materials. For all conditions, it was assumed that the word-monitoring task would prove sufficiently sensitive to reveal difficulties in the on-line processing of late target words in complex sentences. The task was also assumed to be sufficiently sensitive to show that the storage demands of complex sentences exhaust the limited-pool of resources also used by the SAS to supervise the coordination of processing between the visual and the auditory information streams. This was expected to produce a significant interaction between complexity, position and multimodality for the response time measure in the visual-based conditions. None of these predictions was supported by the experimental results: neither did they produce a main effect of complexity, nor a complexity interaction with any of the other factors for the word-monitoring task in both modality-based conditions.

Two explanations can account for these results:
1. Lack of experimental control: The word-monitoring task required a response to a unique target in each sentence. It was expected that subjects would need to read and process sentences whilst scrutinising them for the targets, thus the word-monitoring task would reveal on-line fluctuations of resources. However, since the visual materials were available from the onset of the sentence presentation, subjects were able to locate the visual target before it appeared in the auditory channel. Emphasising that targets should be responded to as soon as recognised and asking subjects to ignore the auditory channel in the MMVA condition undoubtedly contributed to this strategy. These instructions were intended to reveal processing times of early target words and to pinpoint the source of facilitation or interference in multimodal processing. However, they ultimately lead to a de-sensitisation of the response time measure for the complexity factor.
2. Manipulation of complexity: Target words were placed after the processing breakdown point in complex sentences. It was assumed that the attempt to maintain the complex syntactic structure (whose processing load is greater than the threshold-value *K*) in WM would exhaust capacity so as to slow lexical-access and thus, the monitoring response for late-complex target words. This manipulation was strong enough to yield a main effect of complexity for the comprehension measure. However, it is suggested that this manipulation might have been too extreme.

This is supported by the pattern of the response times found for the auditory-based conditions. Table 5.4 shows that for the speech condition, late-simple targets were recognised on average only 81 ms faster than the late-complex targets. Individual results show that the trend is highly inconsistent. A visual scanning for the target cannot explain this pattern of results. It seems that the attentional systems activated the lexical-access system in an equal manner in both the simple and the complex conditions. This might have contributed to the failure to find on-line fluctuations of resources at late-positions in complex sentences. The residual amount of resources failed the syntactic and the semantic systems in the complex condition, as demonstrated by the chance level of performance in the speech and the visual conditions (see Table 5.14). It is possible that using a less complex structure (whose processing load is smaller than, or equal to the threshold-value *K*), would allow the attentional systems to scale the allocation of resources between the lexical-access system and the syntactic and semantic systems in a flexible manner according to their actual needs. This might produce an effect of complexity and an interaction between complexity and position for both modality-based conditions, given the appropriate control of the problems mentioned in (1).

An alternative measure of control could involve placing the target words after the breakdown point, determined by Gibson's NPH principle and before the breakdown point, determined by Gibson's LRNH principle (see Chapter 2). At earlier positions, the competition of resources between the lexical-access system and the syntactic and semantic systems might be more apparent than at the post processing-breakdown position.

In summary, the results of the word-monitoring task do not validate the predictions of the MMUM. The storage demands of partial computational products did not result in a slower identification of late target words in complex sentences. This failure, in itself, does not prove that these predictions are false, due to the reasons just described.

5.5.2 Serial position

The word-monitoring task was successful however, in revealing information regarding the differences between the visual and the auditory modalities in this experiment. An examination of the result found for the position hypothesis in the RT' measure, reveals that late targets were responded to faster (rather than slower) than early target words (see Table 5.13). As suggested earlier, this comparison was justified only for the auditory-based conditions. The visual-based conditions only allowed an examination of response times taken from the onset of the sentence (rather than from the onset of the target). This result contradicts the suggested interpretation of Just & Carpenter's capacity theory. It was believed that the need to maintain intermediate products in WM would slow down almost all computational processes towards the end of the sentence. The inverse result found in this experiment required a further investigation of the speech-processing literature. This investigation aimed to find out whether the facilitation is due to a non-linguistic serial predictability effect or whether it reflects a pure linguistic facilitation of sentential context. Marslen-Wilson & Tyler (1980) conducted a series of experiments using the word-monitoring technique in speech processing. Using a better distribution of target words throughout the sentence, three different monitoring tasks and three

types of prose, these researchers were able to establish that syntactic and semantic cues are built up within the sentence, to create an interpretative framework for word recognition. The possibility that subjects' readiness to respond was simply increased the later that the target occurred in the sentence can be rejected: context facilitates word recognition throughout sentence processing.

Moreover, the overall pattern of results found for the position hypotheses, suggests that some fundamental differences between the two modality-based conditions might have been evident in this experiment. For all conditions, when response times were taken from the onset of the sentence, early target words were recognised faster than late target words. This finding is trivial for the auditory-based conditions, since speech must be perceived serially. For the visual-based conditions, this also reveals some extent of seriality. However, the comprehension rate measure reveals a different pattern of results. Overall, an early response did not improve comprehension rate in this experiment, as indicated by the lack of main effect of position. However, the significant interaction between modality and position demonstrates that the modality-based conditions differed in this respect (see Figure 5.2). As expected, in the auditory-based conditions, an early response leaves more resources for the comprehension of the sentence. On the other hand, in the visual-based conditions, comprehension rates were not affected by the timing of the execution of response, as indicated by the absence of the simple main effect of position for the visual-based conditions. Finally, the absence of a significant interaction between modality, multimodality and position suggests that the unimodal and the multimodal conditions did not differ in this respect in each modality-based condition (see Table 5.10 and Figure 5.3).

It is suggested that the use of regressive eye-movements contributed to the absence of an effect of position in the visual-based conditions. In spite of the assumptions that (i) subjects will read and process the sentences whilst scrutinising them for target words and that (ii) presenting a unique target in each sentence will minimise the use of regressive eye-movements, subjects treated this dual-task experiment as two separate tasks. The requirement for speedy responses to the target words prioritised the execution of the monitoring response and subjects located the visual targets before they appeared in the auditory channel. However, it is suggested that this did not eliminate the use of regressive eye-movements to earlier parts of the text in order to maximise its comprehension. The two processes did not necessarily operate in a clear order. Specifically, it is very likely that response execution for early target words preceded the use of eye-regressions. However, for late target words, regressive eye-movements might have also taken place prior to the execution of the monitoring response. In other words, the requirement for speedy responses to the target words prioritised the execution of the monitoring response, but without an explicit eye-tracking device one cannot exclude the theoretical possibility that regressive eye-movements were used in late-target sentences. Presentation rate in this experiment conformed with a normal reading rate (the duration of presentation of the visual text was fixed to equate with a usual reading speed of 204 words per minute). Also, the visual display of text could not be terminated early by the subject in order to advance more quickly to the true/false question (see Method, section 5.3.1). Thus, although this rate may have been slightly high for the right-branching and the doubly-embedded structures used, it is

reasonable to assume that it enabled subjects to adapt their reading patterns to the main requirements of the two sub-tasks.

Having explained the possible contribution of regressive eye-movements to absence of the effect of position in the visual-based conditions, this cannot solely account for the pattern of results in the auditory-based conditions. The auditory-based conditions, for which the effect of position was identified, also involve a visual presentation in the multimodal condition. If regressive eye-movements were the only factor to account for this pattern of results, their use would have been expected to conceal the effect of position in the auditory-based conditions. A complementary explanation can account for this pattern of results. It involves an informal comparison between the response times found for late-position targets in both the auditory-based and the visual-based conditions (see Table 5.7), and the average presentation time of late-position sentences in all presentation conditions. It appears that in the auditory-based conditions, late targets were responded to 4017 ms measured from the onset of sentence presentation. The average presentation time of these sentences was 4028 ms. On the other hand, in the visual-based conditions, late targets were responded to in 3026 ms measured from the onset of sentence presentation. The average presentation time of these sentences was 4065 ms thus allowing 1039 ms free of target searching. It is highly possible that this "target free" second allowed subjects to reconstruct the assignment of thematic roles to the three lexical noun-phrases and their non-lexical operators, blurring the effect of position for the visual-based conditions. On the other hand, in the auditory-based conditions, late-position targets were responded to at the offset of the sentence. The assignment of thematic roles in these sentences must have been conducted while expecting the appearance of the target. It is not suggested that in the MMAA condition, subjects ignored the external representation of the visual sentence. As will soon be explained, they must have attended the visual channel in this experiment, and probably used regressive eye-movements to earlier parts of the text in order to maximise its comprehension. However, responding to late targets at the offset of the sentence did not enable the use of the visual text to conceal the effect of position in this condition.

Accounting for the absence of the expected interaction between complexity and position for the comprehension rate measure requires a closer look at the values of the complexity, position and modality data (see Table 5.8 and Figure 5.1). In the visual-based conditions, the "target free" second allowed subjects to reconstruct the assignment of thematic roles, blurring the effect of position for both levels of complexity. This strategy cannot however fully account for the absence of a significant interaction between complexity and position, since the data of the auditory-based conditions does not support the prediction either. In the auditory-based conditions, an early response leaves more resources for comprehension regardless of the sentence complexity. (Recall, the execution of an early response was assumed to release more resources for both storage and on-line computing functions and thus to improve the comprehension of simple and complex sentences in a differential manner. Specifically, the difference between the comprehension of early-complex and late-complex sentences was expected to be larger than the difference between the comprehension of early-simple and late-simple sentences).

The account that the manipulation of word-position was too subtle to yield the required effects can be rejected, since it was powerful enough to yield a simple main effect of position for the comprehension rate measure in the auditory-based conditions. Rather, it is suggested that complex sentences might have been too complex to enable the SAS to scale the allocation of attention between the lexical-access system and the syntactic and semantic systems in the predicted manner[30].

5.5.3 Facilitation and interference in multimodal presentation

Given all the interpretation constraints described, what can be learnt about the on-line fluctuation of resources in processing information presented to both modalities? This will be elaborated next.

The word-monitoring task was intended to demonstrate the transition from a facilitatory synchronous processing to an interfering asynchronous processing in the MMVA condition. This was expected to take the form of a significant interaction between multimodality and position. However, results provide only partial confirmation for this prediction (see Table 5.11 and Figure 5.4). The reduction in user cost, approaching significance for early target words, supports the predictions of the MMUM; it appears that the multimodal contact made by the visual and the auditory information in the lexical-access system may indeed benefit processing of early-position words. By contrast, the interference effect found for late-position targets in the MMVA condition is very small and insignificant. An inspection of individual data in the MMVA condition reveals that contrary to expectations, subjects did not attempt to synchronise between the two modalities. For almost half of the subjects, the average word identification time was lower than the average target-onset in the auditory channel.

In addition, the pattern of results found for the monitoring times in the auditory-based conditions is incompatible with the experimental predictions. Results failed to confirm the expected facilitatory effect of multimodality for the auditory-based conditions. It is suggested that the absence of the effect is an experimental artifact that can be clearly demonstrated by the significant interaction between multimodality and position found for the auditory-based conditions. The monitoring times indicate a reduction in user cost for late-position targets but an increase in user cost for early-position targets presented in the MMAA condition (see Tables 5.12 and 5.13, Figures 5.5 and 5.6). This surprising result cannot be successfully accounted for by the MMUM. Since multimodality improves comprehension in the auditory-based conditions, it is clear that subjects processed the visual channel (see Figure 5.8). Moreover, the multimodal interference effect in recognising present catch trials and the absence of a main effect of modality in the multimodal catch trials support this proposition (see Appendix B, Tables B.1 and B.3). The reduction in user cost found for late target words could be explained by a cross-modal priming from the visual to the auditory modality. However, can the MMUM explain the cost increase found for early-position targets?

---

[30] Otherwise, this suggests a serious problem for the capacity theory that claims that sentence length and sentence complexity draw upon the same limited pool of resources.

The account of a cross-modal negative priming can be rejected. Such an account would suggest that when the auditory target word was encountered, subjects were attending to a different visual word, resulting in interference. As suggested by Hanson (1981), unattended, incompatible (single) words do not tend to influence responses at the decision level in multimodal presentations. Moreover, this is consistent with the absence of interference for late targets in the MMVA condition, for which an asynchronous processing was observed.

A better explanation of the cost increase found for early-position targets in the MMAA condition involves an informal comparison between the response times found for early-position targets in both unimodal conditions. Tables 5.11 and 5.12 demonstrate that the response times in the speech condition were 479 ms faster than those found in the visual text condition (1235 ms vs. 1714 ms). If, in the MMAA condition, subjects actively suppressed the "unattended" visual display, its delayed perception may have increased response times to the attended auditory targets. It is not suggested that this is always the case in cross-modal priming from the unattended visual to the attended auditory modality. On the contrary, using the semantic judgement paradigm, Hanson (1981) found that whether presented auditorily or visually, unattended identical and category related words facilitate semantic judgements of the attended word. The same explanation accounts for the facilitation observed in the visual-based conditions: when the (slower) visual modality was attended, the spoken (unattended) targets could successfully prime the visual targets at the early sentential position. The multimodal activation of the lexical-access system might not have been simultaneous, but the (approaching significance) redundancy gain of 238 ms supports the account of a cross-modal priming.

The comprehension measure provides additional evidence of possible strategies in this experiment. First, consider the difference between comprehension rates in the unimodal speech condition (57%) and the unimodal visual text condition (66%), shown in Table 5.15 and Figure 5.8. The low comprehension rates found for the speech condition are particularly striking. These comprehension rates may have been affected by the minimal intonation and prosody included and by the (higher than average) speech rate used in this experiment (see Method, section 5.3.1). In spite of the conceivable contribution of these factors, it is suggested that it is the use of regressive eye-movements in the durable visual text condition which is the main key for understanding the differences between processing speech and visual text in this study. The clearer representations provided by the visual text cannot sufficiently account for these differences, as will soon be explained. As proposed earlier, information in the speech condition must be processed in real-time. If processing of the information cannot be completed immediately, then the user must maintain their primary representations for later processing. The level of complexity of the spoken sentences and the characteristics of the speech itself must have made the immediate processing of the spoken information very difficult in this study. On the other hand, the characteristics of the visual condition enabled a more flexible mode of processing. As noted earlier, presentation rate in the visual condition supported a normal reading rate of the static visual text, making likely the idea of regressive eye-movements to earlier parts of the text. Furthermore, the absence of a main effect of position, found for the comprehension rates in the visual-based conditions was partially explained by the use of regressive eye-movements in the visual-

based conditions. As suggested earlier, this finding suggests that regressive eye-movements operated in the visual-based conditions for both early- and late-target sentences. Therefore, it is proposed that the external representations of the durable visual text condition may have encouraged subjects to allocate resources to sentence computation rather than to storage of intermediate computational products. This de-allocation of resources has probably induced a temporary "forgetting" by displacement that was partially recovered by the recollection of the absent information using regressive eye-movements.

The triple interaction between complexity, modality and multimodality defines the extent of facilitation made by the availability of the visual representations for regressive eye-movements in both complexity conditions. The symmetric pattern of results found for the complex sentences indicates that the two unimodal conditions yielded a similar, chance level of performance. This finding is very important. First, it implies that the clear representations provided by the visual text were not superior to the more ambiguous representations provided by the speech in facilitating the comprehension of complex sentences. It also implies that consistent with the predictions of the MMUM outlined in Chapter 4, the re-collection of absent information from the external representation in the visual condition is equal to scavenging residual mental representations in the speech conditions when processing complex sentences. On the other hand, for the simple sentence structure, comprehension rates in the visual text condition were 12% higher than those found in the speech condition (see Table 5.14 and Figure 5.7). This implies that the clear visual representations might have facilitated the comprehension of simple sentences in comparison to the speech representations, but more importantly, that the re-collection of absent information from the external representation is superior to scavenging residual mental representations only while processing simple sentences. Because the word-monitoring task did not minimise the use of regressive eye-movements for late-position visual sentences (as had been predicted for this experiment), this result reflects a facilitation effect for both early- and late-target visual sentences relative to the speech condition (see Tables 5.5 and 5.6). In conclusion, speech processing and visual text processing do not involve the same proportion of storage and on-line computing, given the visual sentence is fully available for further processing[31]. Whereas this differential allocation of resources proves useful while processing simple visual sentences, it does not improve the comprehension of complex visual sentences. The differences in the allocation of resources between storage and on-line computing that were made in the unimodal and the multimodal conditions in this experiment will be specified next.

Figure 5.7 demonstrates the relationship between complexity and multimodality in both modality-based conditions. Contrary to the experimental predictions, the comprehension of both simple and complex sentences was facilitated by the added visual text in the auditory-based conditions. On the

[31] The claim that the two unimodal conditions involve the same proportion of storage and on-line computing and that the secondary recollection of information is independent and additional to the proportion described is rejected by this account. It seems more likely that the availability of external representations in the static-durable visual text condition enables the user to adapt processing to the task requirements. Since these differ for serial and parallel modes of presentations, it is assumed that speech processing requires a greater allocation of attention to the storage of intermediate computational products than processing of static-durable visual text.

other hand, for the visual-based conditions, multimodality affected the comprehension of simple and complex sentences in a differential manner: the added speech did not affect the comprehension of simple sentences but significantly *improved* the comprehension of complex sentences. For the simple sentences, it is suggested that the availability of the clear durable representations of the visual text encouraged subjects to allocate resources to the visual sentence computation rather than to storage of intermediate computational products in both modality-based multimodal conditions. This de-allocation of resources probably induced a temporary forgetting by displacement that was partially recovered by the recollection of the absent information using regressive eye-movements. The absence of redundancy gain in comprehending simple sentences in the visual-based conditions suggests that this facilitation simply reflects the availability of the visual text for further processing.

How can this explain the facilitation observed in the complex sentences? If subjects allocated more resources to the visual sentence computation rather than to storage of computational products, how could this have helped them to better comprehend the complex sentences in the multimodal conditions? The similar chance performance found in the two unimodal complex conditions, does not suggest that the clear representations provided by the visual text, or a simple recollection of the absent information can successfully account for the redundancy gain found for multimodal comprehension of complex sentences.

The account according to which the intonation and prosodic cues in the spoken sentences assisted the comprehension of complex sentences in the two multimodal conditions can be rejected. The recorded sentences were spoken with the minimum intonation possible. In addition, pauses were edited out to eliminate any reminiscence of prosody. The poor performance observed in the speech condition suggests that this control was sufficient and that intonation and prosody cannot account for the multimodal facilitation observed for complex sentences. In addition, the account that the phonological traces of the complex sentences assisted the interpretation of the complex sentences off-line, after the offset of the sentences, can be also rejected. The MMUM suggests that these traces decay rapidly and could not help the delayed assignment of thematic roles in the three presentation conditions involving speech. The chance performance observed for complex sentences in the speech condition does not suggest that phonological traces of the spoken message facilitated the interpretation of the complex sentences.

Two possible explanations are proposed to explain the multimodal facilitation in comprehending complex sentences. The first explanation implies a special facilitation of multimodality. For the MMVA condition, it suggests that subjects did not simply process the complex sentences by means of a delayed assignment of thematic roles, based only on the visual text. Rather, the availability of the external representations of the visual text encouraged subjects to allocate resources to the *visual* sentence computation rather than to storage of intermediate computational products. This de-allocation of resources has induced a temporary "forgetting" by displacement of the three lexical noun-phrases. However, the static-durable multimodal presentation enabled subjects to regress back visually to the three lexical noun-phrases while attending to the three successive verbs in the *auditory* channel. Thus, according to this explanation, the assignment of the required thematic roles between

the nouns and the verbs was conducted across modalities. Moreover, the relatively early execution of responses for late targets in the MMVA condition enabled subjects to utilise the "target free" second to assign the thematic roles between the nouns and the verbs across modalities with no interference.

The same pattern of performance could also be responsible for the multimodal facilitation observed for the complex sentences in the MMAA condition. The redundancy gain found for the response times of late target words was explained earlier by a cross-modal priming from the (unattended) visual modality to the (attended) auditory modality. Also, the redundancy cost observed for early-position targets was claimed to result from an active suppression of the unattended visual targets. This may imply that in the MMAA complex condition, the availability of the external representations of the visual text encouraged subjects to attend the *visual* channel rather than the auditory channel. The same account suggested for the MMVA condition can therefore be applied for the processing of complex sentences in the MMAA condition with one exception: for the *late-position* complex sentences, the cross-modal assignment of thematic roles must have been conducted while expecting the appearance of the target in the auditory channel. The absence of a four-way interaction between multimodality, modality, complexity and position suggests that this did not affect the extent of multimodal facilitation as compared with the MMVA condition. Contrary to the assumptions of the MMUM, this explanation suggests that the resources used to coordinate the integration of information from the two information channels are independent from the resources used for sentence processing. If both tasks used the same limited pool of resources, one should have observed a multimodal interference for complex sentences in both modality-based conditions.

There is however an alternative explanation for this pattern of results: that the multimodal facilitation found for *complex* sentences is an artificial result. It relies on the finding that the unimodal and the multimodal conditions differed with respect to the ease of performance in the word-monitoring task. The multimodal facilitation effect in recognising target words in the regular trials implies that for both levels of complexity, the word-monitoring task was more difficult in the unimodal conditions relative to the multimodal conditions (see Appendix B, Table B.1). If subjects had to allocate more resources to the word-monitoring task in the unimodal conditions, then fewer resources were available for the comprehension task in these conditions. This shortage of resources might have affected the comprehension of simple and complex sentences in a differential manner: whereas it did not affect the comprehension of the simple sentences, it might have impaired the comprehension of complex sentences in the unimodal conditions relative to the multimodal conditions.

The inspection of individual results, reported earlier, provides support for this explanation. This inspection revealed that in each of the unimodal conditions, there were two target words which approximately one-third of all subjects mistook for catch trials. If these particular target words were difficult to recognise, this might have impaired the comprehension of the complex sentences in which they appeared regardless of the correctness of subjects' responses to the actual target words. No indication of specific difficulties was found for the multimodal conditions.

One objection could be made against this alternative account. In order to comply with instructions, subjects were to ignore the auditory channel in the MMVA condition and the visual channel in the MMAA condition. The need to ignore specific channels in each of the multimodal conditions implies a qualitatively different kind of difficulty: if subjects had to actively suppress attention to specific channels in each of the multimodal conditions, then *less* resources were available for sentence comprehension in the multimodal conditions than in the unimodal conditions. However, the only indication for a higher processing cost due to an active suppression of specific modalities in the multimodal conditions is provided by the response times data for early-position targets in the MMAA condition. The overall pattern of results clearly implies that subjects disobeyed the requirement to ignore specific channels in this experiment. Specifically, the catch trials' data indicates an allocation of attention to the prohibited channels as the percentage of missing an existing catch trial was significantly larger for the multimodal conditions relative to the unimodal conditions (see Appendix B, Table B.1). This simply implies that subjects did not attend the required channels in the multimodal regular-trials, and does not undermine the alternative explanation provided for the multimodal facilitation in comprehending complex sentences.

Overall, the pattern of results necessitates a serious modification of the MMUM. In the concluding section, I will try to elaborate on the required modifications and on further experiments needed to clarify the validity of the results reported throughout this section.

5.6 Conclusions

In this study, language processing demands (storage and on-line computing demands) were assumed to draw upon the limited pool of resources used by the SAS for its coordination of both the sentence processing and the word-monitoring tasks. The experimental manipulation was only partially capable of validating the assumed relationships between language processing demands and word-monitoring demands. Central to the MMUM, the experimental manipulation failed to reveal on-line processing difficulties of late target words in complex sentences: neither a main effect of complexity, nor a complexity interaction with position for the response time measure was observed in any of the presentation conditions. It is suggested that the availability of the visual text from the onset of sentence presentation, led to a de-sensitisation of the response time measure for the complexity factor. It is also suggested that the manipulation of complexity might have been too strong to identify any of these effects for the response time measure, as complexity did not affect monitoring times of late target words in the speech condition. Finally, since neither the visual-based conditions nor the auditory-based conditions showed an interaction between complexity and position for the comprehension measure, it is suggested that the complex sentences might have been too complex to enable the SAS to scale the allocation of attention between the lexical-access system and the syntactic and semantic systems in the predicted manner.

Most importantly, the experimental results challenge the assumed relationship between language processing demands and synchronisation between modalities suggested by the MMUM. As described above, the significant interaction between multimodality and position found in both modality-based

conditions for the response time measure indicates that subjects did not synchronise between the two information channels in a linear fashion (i.e., by bringing "out of phase" stimuli into phase during processing). It is suggested that the priority given to the word-monitoring task over the comprehension task cannot invalidate the assumption that under natural processing conditions, some form of synchronised processing does take place. However, the only cost observed for this asynchronous behaviour was in the word monitoring times of early-position targets in the auditory-based conditions. It is suggested that this is an experimental artifact, resulting from subjects' attempts to suppress the unattended visual target. In support, the comprehension values reveal that, whereas for simple sentences multimodality improves performance as compared to speech presentation, for complex sentences multimodality improves performance for both modality-based conditions. It is a particularly unexpected and surprising result that speech added to visual text will improve comprehension when that text is complex.

Two possible explanations were raised to account for the multimodal facilitation in comprehending complex sentences. According to the first explanation, a fully redundant multimodal presentation of long-complex sentences does not impose a higher verbal memory load on the user than a single modality presentation of such sentences. Users are able to fully utilise the affordance of the visual presentation for a further recollection of information. They allocate more resources to sentence computation rather than to storage of intermediate and final products that impose an unacceptable processing load on the syntactic parser. Absent information can later be recollected using regressive eye-movements. Significantly, the further recollection of the absent information is beneficial only in a fully redundant multimodal presentation that enables a switch of attention between modalities, so as to assign thematic roles across them. According to this explanation, the SAS is capable of adapting to modality-specific requirements in a flexible manner: its allocation of resources between the different sub-systems optimises the achievement of the task while minimising the cost of the global processing. The SAS is capable of monitoring the performance of the language processing system and to control the competition between different sub-systems across modalities. Finally, this explanation implies that the resources used by the SAS to coordinate the integration of information across modalities are independent from the resources used for sentence processing.

The second explanation has less severe consequences for the MMUM. It suggests that the multimodal facilitation in comprehending complex sentences is an artificial finding that arises due to methodological limitations in the design of this experiment. It relies on the finding that, for both levels of complexity, the word-monitoring task was significantly more difficult in the unimodal conditions relative to the multimodal conditions. If subjects had to allocate more resources to the word-monitoring task in the unimodal conditions, then fewer resources were available for the comprehension task in these conditions. This shortage of resources might have affected the comprehension of simple and complex sentences in a differential manner: whereas it did not affect the comprehension of the simple sentences, it might have impaired the comprehension of complex sentences in the unimodal conditions relative to the multimodal conditions. According to this explanation, the predictions of the MMUM remain untested. Excluding the word-monitoring task

would be expected to reveal that the efficiency of dividing attention between aural and visual stimuli disappears when processing complex sentences.

The next experiment will attempt to clarify the validity of these two possible explanations. Using a simpler experimental design, the experiment aims to assess the role of *durability* in multimodal processing of sentences that vary in their syntactic complexity. If the first explanation is valid and users are able to assign thematic roles across modalities, then for a durable visual presentation, one would expect to obtain a multimodal facilitation effect in comprehending complex sentences. On the other hand, if the second explanation is valid and the difficulties in performing the word-monitoring task in the unimodal conditions account for the results of this experiment, one would expect to observe the opposite effect as predicted by the MMUM. Complex sentences presented to both modalities would then yield lower comprehension rates relative to a unimodal presentation of such sentences.

Chapter 6

Experiment 2: the effect of the durability of the visual text on comprehending simple and complex sentences presented to both modalities

Experiment 2 consisted of two experiments. The first experiment (2a) examined two related questions: (i) what the effect of multimodality is on user cost, given variations in syntactic complexity and (ii) what the added effect of a durable visual text is on user cost, given variations in syntactic complexity. As in experiment 1, syntactic complexity was systematically varied based on Gibson's complexity metric. The durability of the visual text was systematically varied using two dynamic presentation techniques: a dynamic-transient format in which words were presented one by one on the centre of the screen and a dynamic-durable format in which words accumulated on the screen to form a sentence. Multimodal presentation consisted of presenting the two visual text conditions with redundant coupled speech. The results of experiment 2a indicate that, regardless of sentence complexity, presentational durability and subjects' verbal WM capacity, multimodality impairs sentence comprehension. The results also demonstrate higher comprehension rates of sentences presented in the dynamic-transient formats relative to those presented in the dynamic-durable formats. As the low comprehension rates of sentences presented in the dynamic-durable formats could not be fully accounted for by the MMUM, it was decided to investigate other potential variables that might have played a role in this experiment. The results of various analyses imply that in the dynamic-durable conditions, the use of line wrapping with a 15 inch monitor impaired subjects' ability to predict the location in which words were to appear as more than half of the sentences were presented over two lines. Towards the end of the first line, subjects could not predict whether a short word would appear at the end of the same line, or a longer word at the beginning of the next line. Greater confusion took place in the multimodal conditions: it is suggested that when subjects misjudged the location in which the next word was to appear, the normal continuation of the spoken sentence might have impaired their ability to refocus on the "leading edge" of the visual sentence.

Separate analyses were conducted for the single- and the double-line sentences and their results reveal different patterns of interference. For instance, the analysis conducted for the double-line sentences yielded a stronger speech interference effect for the durable conditions than for the

transient conditions. Furthermore, speech impaired comprehension in the durable conditions regardless of sentence complexity. For the single-line sentences, the pattern of results differed substantially. For these sentences, a significant interaction between durability, multimodality and complexity was identified: whereas for simple sentences, the addition of speech had no effect on comprehension rates, multimodality impaired the comprehension of complex sentences. Further analyses revealed that the source of the cost found for complex sentences was in the transient conditions. For the dynamic-durable mode of presentation, the addition of speech had no effect on comprehension rates of complex sentences. Due to the small number of single-line sentences in this experiment, this result was found unreliable. It was decided to run the experiment again using a 17 inch monitor.

Experiment 2b was a replication of experiment 2a. Using a 17 inch monitor, each sentence in the dynamic-durable conditions was presented in a single line. The results of experiment 2b indicate that multimodality impairs the comprehension of complex sentences. Specifically, in the dynamic-durable condition, speech interfered with the comprehension of complex sentences more than it did in the dynamic-transient condition. In addition, the results show that regardless of the complexity of the sentences, high capacity subjects are more resistant to multimodal interference than low capacity subjects. These findings contradict the results of experiment 1 and provide partial support to the original predictions of the MMUM.

6.1 Experiment 2a

6.1.1 Introduction

This chapter reports a study that aims to examine whether the durability of a dynamic visual text affects user cost during multimodal processing of sentences that vary in their syntactic complexity. With the aim of informing the design of systems wherein users can successfully comprehend various texts with a minimum of processing cost, this study incorporates the same principles for the assessment of user cost mentioned in the previous chapters. The assessment of user cost takes into account the linguistic complexity of the presented materials, the memory demands incurred by their multimodal presentation type and also the verbal WM capacity of the user.

Experiment 2a relies on a simpler experimental design than experiment 1. In experiment 1, the assessment of user cost consisted of the dual-task paradigm: as well as monitoring for target words, subjects had to perform a comprehension task for each sentence. The word-monitoring task served as an on-line measure of attention for processing sentences presented to both modalities and the comprehension task ensured that the sentences were properly processed. It is suggested that the word-monitoring task might have confounded the experimental results reported in the previous chapter. Specifically, this proposal refers to the following finding: whereas for simple sentences multimodal presentation improved performance relative to the speech condition, for syntactically complex sentences multimodal presentation improved comprehension relative to both the speech and the

static-durable visual text conditions. Two possible explanations were raised to account for the multimodal facilitation in comprehending complex sentences.
i) The durability account: users return to the durable visual presentation for a further recollection of information. They allocate more resources to sentence computation rather than to storage of intermediate and final products that impose an unacceptable processing load on the syntactic parser. The absent information can later be recollected using regressive eye-movements. Significantly, the further recollection of the absent information is beneficial only in a durable multimodal presentation that enables a switch of attention between modalities, so as to assign thematic roles across them.
ii) The methodological account: the multimodal facilitation in comprehending complex sentences is an artificial finding that arises due to the different levels of difficulty of the word-monitoring task in the unimodal and the multimodal conditions. It relies on the finding that for both levels of complexity, the word-monitoring task was significantly more difficult in the unimodal conditions relative to the multimodal conditions. If subjects had to allocate more resources to the word-monitoring task in the unimodal conditions, then fewer resources were available for the comprehension task in these conditions. This shortage of resources might have affected the comprehension of simple and complex sentences in a differential manner: whereas it did not affect the comprehension of the simple sentences, it might have impaired the comprehension of complex sentences in the unimodal conditions relative to the multimodal conditions.

These two explanations differ in their consequences for the validity of the MMUM. According to the durability account, the MMUM consists of false assumptions. The MMUM assumes that the SAS relies upon the same limited pool of resources used for sentence processing for its supervision functions. Thus, increasing sentence complexity imposes demands not only on the resources that are used by the language processing system, but also on the resources used by the SAS. Consequently, the ability of the SAS to supervise the coordination of processing between modalities (by bringing "out of phase" stimuli into phase) will be impaired. In contrast, the durability account implies that the resources used by the SAS to supervise the coordination of information between modalities are independent of the resources used for sentence processing. Given a durable visual presentation, the SAS is capable of monitoring the performance of the language processing system and to supervise the coordination of information between modalities regardless of sentence complexity. Finally, the methodological explanation suggests that the predictions of the MMUM remain untested. With the removal of the word-monitoring task, the efficiency of dividing attention between aural and visual stimuli might cease to exist when processing complex sentences.

A simpler design is therefore required to decide between the two explanations. In this experiment, the assessment of the durability factor was conducted under less restrictive processing conditions relative to those of experiment 1: both the word-monitoring task and the instruction to attend a specific modality in the multimodal presentation conditions were removed. By giving up the word-monitoring task, an important source of information regarding the on-line fluctuation of processing-resources would be lost but nevertheless, an interpretable set of data would be gained. With the single requirement to comprehend sentences varying in syntactic complexity, a systematic manipulation of

presentational durability will provide a more accurate account of processing cost of sentences presented to both modalities.

The manipulation of presentational durability consisted of two dynamic presentation techniques: (i) the dynamic-transient visual presentation whereby visual text had a transient form as words were presented one at a time at a fixed location on the screen, and (ii) the dynamic-durable visual presentation where the visual text had a serial-durable form: words appeared one at a time and accumulated on the screen to form a sentence. According to the MMUM, these two presentation techniques do not involve the same memory demands. As noted earlier, these demands do not refer to the constraints specified by Gibson and thus, to the syntactic complexity value of the sentence. They refer to the different storage and computational demands that are required by transient and durable presentation techniques for successful processing of a given sentence. Verbal information presented by dynamic-transient media must be processed in real time; if the information cannot be computed immediately, the user must maintain their primary representations for later processing. This suggests that dynamic-transient media imposes high memory demands when the sentences presented are relatively long or complex. The status of a dynamic-durable visual presentation is slightly different: a dynamic-durable visual text forces the pace of reading but nevertheless provides a partial form of durability. Significantly for the purpose of this research, this presentation technique allows the use of regressive eye-movements. However, since dynamic media attracts attention automatically, the MMUM proposes that regressive eye-movements might be more difficult to carry out under dynamic-durable visual presentation than under static-durable visual presentation. Moreover, given a fast rate of presentation, regressive eye-movements must be conducted at the expense of processing a later sentential component. Despite the limitations of the dynamic-durable visual technique, its contrast with the dynamic-transient visual technique should enable an assessment of the pure role of presentational durability in sentence processing. These two dynamic visual presentation techniques enable control of presentation rate and guarantee equal intelligibility for both visual contents. Moreover, whereas the addition of speech to the dynamic-transient visual condition enables an assessment of the effect of multimodality in itself on sentence processing, the addition of speech to the dynamic-durable visual condition creates a multimodal condition that enables the user to physically recollect early sentential components while simultaneously attending to the spoken continuation of the sentence. This is precisely the pattern of behaviour suggested by the durability account for the multimodal facilitation in comprehending complex sentences that was observed in experiment 1.

In this study, multimodal presentation consisted of presenting the two dynamic visual text conditions with additional speech in a coupled manner, so that the outputs of the two modalities occurred at the same time. The word-monitoring data of experiment 1 clearly demonstrates that subjects did not attempt to synchronise the processing of the visual words with the spoken words. However, under natural processing conditions, without the priority given to the word-monitoring task over the comprehension task, some form of synchronised processing might take place. Experiment 2a approaches the investigation of synchronous processing from a different angle, as both multimodal conditions presented the visual and the auditory words together at the same rate. In this, the study

avoids the synchronisation problem specified by the MMUM for a static-durable multimodal presentation: whereas for early sentential components, the lexical sub-system is accessed with redundant information, late sentential components will not necessarily make contact on a redundant multimodal basis without additional control of reading pace. According to the MMUM, when the visual and the auditory stimuli are presented together at the same rate, more resources can be allocated to the language system than when presentation is unimodal. The multimodal contact in the lexical sub-system enables processing to feed the a-modal sub-systems with consistent representations of the verified words. Given the assumption that a common pool of resources serves all language processing sub-systems, this multimodal activation implies that more resources are available to the syntactic and the semantic sub-systems. Furthermore, consistent representations maintained by the phonological sub-system enable post-interpretative utilisation of phonological information. Thus, in providing two coupled multimodal presentation techniques, this study enables an assessment of the role of durability of the visual presentation with all other factors aiming to optimise sentence processing.

6.1.2 Experimental hypotheses

This section provides detailed predictions for the comprehension rate measure in this experiment. Most of the predictions do not refer to the time it takes subjects to respond to each comprehension statement. Except the sentence complexity factor, none of the other experimental factors was assumed to affect the time it takes to comprehend each statement. Response times were simply collected as a measure of control against unexpected trade-offs between the speed of responses and their accuracy.

Sentence complexity

The MMUM suggests that the excessive storage demands imposed by the need to maintain several unassigned thematic roles in memory will lead to a breakdown during processing of the doubly-embedded sentence structure (c.f., Gibson, 1991). Specifically, the syntactic parser will fail to maintain the syntactic structure whose processing load is greater than the threshold-value *K*. The model also suggests that the semantic sub-system will face difficulties in interpreting the semantic relationship between constituents, such as the role relationship of the verb, properties of objects and temporal relationships between objects or events. Since these cannot be interpreted on-line, the model assumes that the phonological sub-system will try to maintain word-order information active to enable post-interpretative utilisation of phonological representations. This is bound to fail since referring to the phonological form and the order of lexical items in the sentence can only reinitiate the parsing process and cannot yield meaning directly (Waters et al., 1987). The failure to maintain the syntactic structure of the doubly-embedded sentences in WM is therefore expected to yield lower comprehension rates and higher response times than those found for the right-branching sentences in all presentation conditions.

Verbal WM capacity and its relationships with durability, multimodality and sentence complexity

According to the capacity theory, individual differences in verbal WM capacity are mostly apparent when a linguistic task imposes an excessive load on the users. Following this account, the MMUM hypothesises that individual differences in verbal WM capacity will affect performance when the combination of the multimodal presentation technique and the linguistic complexity of the presented materials imposes a high processing load on the users. The higher the processing load, the more apparent the difference will be.

The MMUM predicts a significant effect of verbal WM capacity. Overall, high span subjects are expected to show higher comprehension rates than low span subjects for both simple and complex sentences in all presentation conditions. In addition, it is proposed that the span variable will significantly interact with the durability and multimodality variables only for complex sentences. The expected pattern of results for simple sentences is described next.

Simple sentences: The MMUM proposes that regressive eye-movements will be difficult for all subjects to carry out when right-branching sentences are presented in dynamic-durable visual form. Subjects are not expected to successfully recollect previously processed information without losing later sentential components. Thus, performance differences between span subjects in the dynamic-durable visual presentation are expected to be similar to those in the dynamic-transient visual condition (see Chapter 4, Figures 4.4 and 4.5).

A slightly different pattern of results is predicted for the multimodal conditions. According to the MMUM, the addition of redundant coupled speech is expected to reduce user cost slightly for both durable and transient forms of presentation. As noted in chapters 3 and 4, such a coupled presentation of visual and auditory words will increase the activation of the lexical-access sub-system. This multimodal activation will optimise information processing in the a-modal sub-systems that are fed by consistent representations of the verified words, resulting in a slightly lower user cost for all subjects. Differences in verbal WM capacity are expected however to slightly affect performance while processing long-simple sentences in the dynamic-durable multimodal condition. According to the suggested framework, a dynamic-durable multimodal presentation of long-simple sentences enables users to attend to the spoken continuation of the sentence while performing regressive eye-movements. The incompatibility of the spoken information with the recollected visual information is expected to produce a local interference effect for low span subjects. The SAS of these users is assumed to have insufficient resources to accommodate both storage and computational demands of long-simple sentences and to maintain successful divided attention between modalities. Since synchronous processing is easy to restore in the dynamic-durable multimodal condition through refocusing attention on the "leading edge" of the visual display, an interference effect is not expected in this condition relative to the dynamic-transient multimodal condition. Thus, for low span subjects, comprehension rates of long-simple sentences are expected to be equal for both multimodal formats (see Chapter 4, Figure 4.4). On the other hand, high span subjects are expected to have sufficient resources to accommodate both storage and computational demands of long-simple sentences and to

maintain successful divided attention between modalities. The performance of these subjects in the dynamic-durable multimodal condition is expected to be higher than their performance in the dynamic-transient multimodal condition (see Chapter 4, Figure 4.5). The expected weak facilitation effect in the dynamic-transient multimodal condition (predicted for all subjects) might conceal this mild interaction. Nevertheless, it is questionable whether the multimodal facilitation in both durability conditions will be strong enough to influence the comprehension rates of long-simple sentences.

Complex sentences: For the doubly-embedded structure, a significant interaction between span, durability and multimodality is forecasted. The MMUM assumes that the use of regressive eye-movements is not expected to assist the interpretation of complex sentences in the dynamic-durable visual condition, as both low and high span subjects will not have sufficient resources to recollect previously processed information without losing later sentential components. In this form of presentation, the syntactic parser will not have sufficient resources available for the delayed assignment of thematic roles required for the comprehension of complex sentences. So, for complex sentences presented visually, the model predicts a main effect of span, no effect of durability and no interaction between span and durability (see Chapter 4, Figures 4.6 and 4.7).

A different pattern of performance is predicted for multimodal presentation of complex sentences. As noted earlier, the dynamic-transient multimodal technique does not enable the user to perform regressive eye-movements, and is therefore more resistant to any breakdown of synchronisation. Since the simultaneous cross-modal activation of the lexical-access sub-system is assumed to be negligible under these circumstances, similar comprehension rates are predicted for both unimodal and multimodal dynamic-transient conditions; higher for high span subjects than for low span subjects (see Chapter 4, Figures 4.6 and 4.7).

On the other hand, the dynamic-durable multimodal technique enables the user to use eye regressions while attending to the spoken continuation of the sentence. Again, this is precisely the pattern of behaviour suggested by the durability account for the multimodal facilitation in comprehending complex sentences that was observed in experiment 1. If this account is valid and the durability of the visual presentation does enable the SAS to supervise the coordination of information between modalities, one would expect to obtain a higher comprehension of doubly-embedded sentences in the dynamic-durable multimodal condition than in the dynamic-durable visual condition. Moreover, the facilitation in comprehension should be larger for high span than for low span subjects.

However, the MMUM predicts the opposite. According to the model, the use of regressive eye-movements while attending to the spoken continuation of an excessively complex sentence will be an unsuccessful strategy. Specifically, the incompatibility of the spoken information with the recollected visual information will produce an interference effect in the cross-modal sub-systems for all users. As a result, conflicting representations will access the a-modal storage space that serves the syntactic and semantic sub-systems. The MMUM assumes that the SAS of all users will not have sufficient resources available to supervise the competition between the language sub-systems for activation and

the coordination of information between modalities. Finally, since low span subjects are assumed to have lower verbal WM capacity than high span subjects, a greater multimodal interference effect is predicted for these subjects in comprehending long-complex sentences delivered by a dynamic-durable presentation (see Chapter 4, Figures 4.6 and 4.7). In summary, for long-complex sentences presented dynamically to both modalities, the model predicts a main effect of span, an interference effect of durability and an interaction between span and durability.

On average, and regardless of users' verbal WM capacity, the MMUM predicts a significant triple interaction between complexity, durability and multimodality. For simple sentences, neither a significant main effect of durability, nor a significant main effect of multimodality is expected. In addition, the two variables are not expected to interact (see Chapter 4, Figure 4.2). Specifically, the model assumes that for both durable and transient forms of presentation, the addition of speech is expected to slightly reduce user cost. As noted earlier, it is questionable whether this reduction of processing cost will be strong enough to influence the comprehension of long-simple sentences. In addition, the MMUM predicts that for both unimodal and multimodal presentation conditions, the durability of the visual text will not affect the comprehension of long-simple sentences. For complex sentences, the model predicts a significant interaction between durability and multimodality (see Chapter 4, Figure 4.3). As suggested earlier, for the dynamic-transient conditions, the simultaneous cross-modal activation of the lexical-access sub-system is assumed to be negligible under the excessive processing demands of the doubly-embedded structure. For these sentences, similar comprehension rates are expected for both unimodal and multimodal dynamic-transient presentation conditions. On the other hand, whereas the comprehension of complex sentences presented in the dynamic-durable visual condition is expected to equal the comprehension in the dynamic-transient visual condition, multimodality is expected to impair the comprehension of complex sentences when these consist of a dynamic-durable visual presentation. Specifically, the incompatibility of the spoken information with the recollected visual information will produce an interference effect at the lexical-access sub-system and hence, at the syntactic and semantic sub-systems. Again, the MMUM assumes that the SAS will not have sufficient resources available to supervise the competition for activation between the language sub-systems and to oversee the coordination of information between modalities.

Overall, this pattern of behaviour will lead to a significant interaction between complexity and multimodality: a slight facilitation of multimodality in the comprehension of long-simple sentences and a weak interference effect of multimodality in the comprehension of long-complex sentences.

6.1.3 Method

Materials and design

The primary stimulus set consisted of 80 pairs of sentences, each containing one right-branching sentence and one doubly-embedded sentence. These sentences were a shorter version of the sentences used in experiment 1, since the adjectives that were used as target words in that experiment were removed.

Subjects were assigned to one of two experimental groups defined by the complexity factor[32]. Each group was presented with the sentences in four presentation conditions created by the durability and the multimodality factors[33]:

1. Dynamic-transient visual text
2. Dynamic-durable visual text
3. Dynamic-transient multimodal presentation
4. Dynamic-durable multimodal presentation

Digital recording and editing of speech was conducted using the SoundEdit™ application (sample rate: 22.05 kHz, sample size: 16 bits), resulting in better sound quality than in experiment 1. The sentences were spoken in a male voice, with minimum intonation and prosody[34]. They varied in length (from 11 to 15 words) and had an average presentation rate of 159 words per minute, slower than the average presentation rate in experiment 1. Presentation duration of each sentence took on average 4990 ms.

In order to achieve a coupled multimodal presentation, the audible duration of each spoken word had to be measured and the word needed to be displayed visually for this duration. For consistency, the same measure was used for the visual presentation conditions. Sentences were recorded and stored as

---

[32] See Chapter 5, footnote 23.

[33] This study maintained the division of presentation conditions into uniform blocks in order to minimise subjects' confusion. It was assumed that performance would be disrupted by changing from one mode to another. Specifically, if the change of mode had been very heavily cued, this cueing would have distracted subjects from the language processing task, but without cueing the uncertainty might have interfered with the task even more profoundly. Having a presentation mode delivered in a uniform block was not believed to be a problem - any potential practice effect within block was assumed to equate across the different presentation conditions.

[34] The recurrent use of sentences with minimal intonation and prosody aimed to maintain the equal predictive value of the visual and the auditory modalities with regard to the syntactic boundaries of the structures used, but also to minimise the use of recall strategies in the simple sentences condition. For the complex sentences condition, the use of minimal intonation and prosody was not expected to affect performance (the use of intonation and prosody may assist comprehension only when the canonical order of the agents and the themes in the sentence can be easily identified).

separate waveform files. These waveforms were edited so that both the simple and the complex versions of each sentence had the same presentation duration. In addition, for each sentence, a marker was placed in the waveform at the locations defining the boundaries of each word. Whenever possible, the marker was placed in the waveform in areas of low signal amplitude, as indicated by auditory and visual inspection. However, in natural speech, word boundaries often do not coincide with areas of low signal amplitude. In these cases, the markers were positioned so as to maximise the intelligibility of the affected words. The resulting *inter-marker times* provided the audible duration of all spoken words and were listed for each sentence. 40 recorded sentence pairs were selected and allocated to the two multimodal presentation conditions that included spoken output: the dynamic-transient multimodal condition and the dynamic-durable multimodal condition.

A visual text version was created for each of the sentences using Palatino 18 points font. Visual sentence pairs were allocated to the four presentation conditions for each complexity group. To create the dynamic-transient visual presentation, separate PICT files were made for each visual word, each file containing one word. Presenting these PICT files in sequence for the duration of their *inter-marker times*, words appear one at a time at a fixed location on the screen. Similarly, to create the dynamic-durable conditions, separate PICT files were made for each visual word, each file containing a word and all other words that precede it in a given sentence. For example, for the first sentential word, the PICT file included one visual word. For the second sentential word, the PICT file included both the first and the second visual words and so on. The file of the last sentential word included the full sentence. Presenting these PICT files in sequence for the duration of their *inter-marker times*, words appear one at a time but accumulate on the screen to form a sentence.

Comprehension statements

Comprehension statements were created using the same principles described in experiment 1.

Implementation

Similarly to experiment 1, the implementation of these materials was made using the SuperLab™ application. For each complexity group, presentation conditions were grouped in four different blocks. Each block consisted of 20 sentences. Sentences were balanced across presentation conditions with respect to their length (on average: 13.2 words) and their pragmatic complexity[35]. In addition, the comprehension statements were balanced across presentation conditions with respect to the number of true/false statements[36] and their internal distribution of the nouns-verb combinations. The presentation of the sentences was randomised within each presentation block.

---

[35] See Chapter 5, footnote 25.

[36] See Chapter 5, footnote 26.

Four experimental versions were created for each complexity group each having a different order of the presentation-conditions blocks to minimise order effects[37]:
Version 1: dynamic-durable visual presentation, dynamic-durable multimodal presentation, dynamic-transient visual presentation and dynamic-transient multimodal presentation.
Version 2: dynamic-transient visual presentation, dynamic-transient multimodal presentation, dynamic-durable visual presentation and dynamic-durable multimodal presentation.
Version 3: dynamic-durable multimodal presentation, dynamic-durable visual presentation, dynamic-transient multimodal presentation and dynamic-transient visual presentation.
Version 4: dynamic-transient multimodal presentation, dynamic-transient visual presentation, dynamic-durable multimodal presentation and dynamic-durable visual presentation.

Apparatus

The experiment was run on a PowerPC 5200/75. Spoken sentences were presented at a comfortable volume through headphones (Vivanco SR 250) and the visual material was presented on a standard Apple display, size 15 inch. Subjects' responses to the comprehension statements and the times of these responses were collected via Apple Desktop Bus™ (ADB) keyboard (see experiment 1).

Procedure

Subjects first read a comprehensive set of instructions, describing the experimental task and the sentence structure they would encounter, illustrated with an example. They were told that their task consisted of judging true or false statements, which would test their understanding of each presented sentence. Subjects were also told that after the sentence had disappeared, they would be presented with a statement, which they would have to judge as true or false (by pressing the "Z" key for a true statement and the "X" key for a false statement). They were told to respond as quickly and accurately as they could and that an alert sound (a "Beep") would be given for an erroneous response. After confirming that the instructions were understood, the practice session started. The practice was divided to the 4 different presentation conditions described above. Subjects were presented with specific instructions for each presentation conditions followed by 12 practice sentences.

Span task

Subjects' verbal WM capacity was measured in a separate session. The same testing procedure was followed as in experiment 1. 18 subjects whose reading spans were 3.5 or higher were classified as high span subjects and 14 subjects whose reading spans were 3.0 or lower were classified as low span subjects.

---

[37] See Chapter 5, footnote 27.

Subjects

32 students and faculty members at University College London were tested and were paid £6 for their participation. English was the first language of all subjects.

6.1.4 Results & Discussion

Subjects' Analysis of the full data set

Comprehension rates

Comprehension rates were collected for each subject and were analysed in a repeated-measure analysis of variance. Complexity and span formed the between-subjects variables while durability and multimodality formed the within-subjects variables[38],[39]. As expected, comprehension rates of simple sentences (87%) were higher than those found for complex sentences (68%), yielding a main effect of complexity ($F(1, 28) = 41.601$; $p < .001$)[40]. However, the overall pattern of results is not

---

[38] An exploration of the comprehension rate data revealed that the sub-conditions created by complexity, durability and multimodality did not exhibit perfectly normal distributions. The assumption of normality was not met for simple sentences in the dynamic-transient visual condition (Shapiro-Wilk (16) = .876; $p < .04$). Furthermore, their distributions varied in shape: 4 sub-conditions were positively skewed and 4 were negatively skewed. The assumption of homogeneity of variance was not met for both the dynamic-transient visual and the dynamic-transient multimodal conditions. Furthermore, the sub-conditions created by span, durability and multimodality did not exhibit perfectly normal distributions. Specifically, the assumption of normality was not met for low span subjects in the dynamic-transient visual condition (Shapiro-Wilk (14) = .854; $p < .03$). An additional exploration used the squared values of the comprehension rate data. Again, it was found that for the sub-conditions created by complexity, durability and multimodality, the assumption of normality was not met for simple sentences in the dynamic-transient visual condition (Shapiro-Wilk (16) = .878; $p < .04$). Their distributions also varied in shape: 6 distributions were positively skewed and 2 were negatively skewed. However, for the sub-conditions created by span, durability and multimodality, the assumption of normality was met for all presentation conditions. Moreover, for the squared values, the assumption of homogeneity of variance was now met for all presentation conditions created by both the complexity, durability and multimodality variables and the span, durability and multimodality variables. It was decided to conduct the analysis of variance over the transformed values of the comprehension rate measure rather than the original values. Note that the reported mean values were converted back to the original units using a square root transformation.

[39] An additional analysis was conducted at the request of the examiners of this dissertation. This analysis included the order in which subjects performed the four presentation conditions as an additional between-subjects variable (version number) to make sure that the experimental results were unaffected by order effects (see Method, section 6.1.3). None of the results reported in this section was affected by the order in which subjects performed the four presentation conditions. Note that the number of subjects in each version is too small to make this a reliable conclusion.

[40] Note that the comprehension accuracy of all sentences was higher than in experiment 1. This is a result of the removal of the monitoring task and the slower presentation rate used in this experiment. Furthermore, a One-

compatible with the predictions made for this experiment. Contrary to the experimental predictions, comprehension rates of high span subjects (78%) were only 3% higher than those found for low span subjects (75%), failing to yield a significant effect of span (F (1, 28) = 2.524). More importantly, in contrast to the specific speech interference effect that was predicted for complex sentences in the dynamic-durable multimodal condition, results demonstrate that regardless of sentence complexity, presentational durability or subjects' verbal WM capacity, the addition of speech impaired comprehension: comprehension rates in the multimodal conditions (75%) were 6% lower than the rates found for the two visual conditions (81%), yielding a significant effect of multimodality (F (1, 28) = 12.431; $p < .01$). In addition, the analysis indicates that for both unimodal and multimodal presentation techniques, the durability of the visual text impaired comprehension of both sentence types: the results demonstrate a significantly higher cognitive cost for the dynamic-durable presentation techniques (75%) relative to the dynamic-transient techniques (80%), (F (1, 28) = 8.749; $p < .01$).

Finally, the analysis yielded a significant 4 way interaction between complexity, durability, multimodality and the span factors (F (1, 28) = 4.756; $p < .05$). Table 6.1 and Figure 6.1 present the mean comprehension rate found for simple sentences. Table 6.2 and Figure 6.2 present the means found for complex sentences·

Table 6.1

Experiment 2a - Full Data. Mean Comprehension Rate (CR) for the Simple Conditions (%): By Span, Durability and Multimodality (Standard Errors in Parentheses)

WM Capacity	Durability by Multimodality			
	dynamic-durable visual	dynamic-durable multimodal	dynamic-transient visual	dynamic-transient multimodal
High	92%	85%	89%	89%
	(2%)	(4%)	(3%)	(4%)
Low	86%	79%	91%	82%
	(3%)	(5%)	(4%)	(5%)
Mean	89%	82%	90%	86%

Sample T-Test revealed that for this study, comprehension rates of complex sentences were higher than chance level in all presentation conditions (t_8 = 2.828-11.701, $p < .05$), except for low span subjects in the dynamic-durable multimodal condition (t_8 = 1.263. $p < .05$).

Figure 6.1

Experiment 2a - Full Data. Mean Comprehension Rate (CR) for the Simple Conditions (%): By Span, Durability and Multimodality

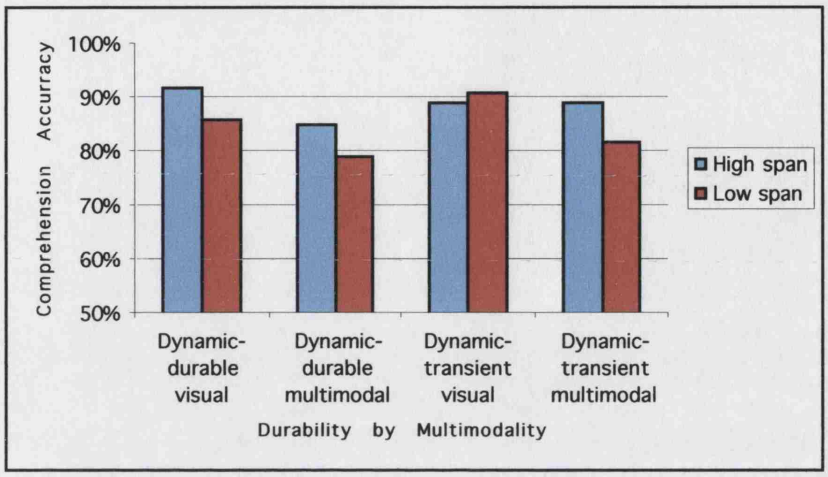


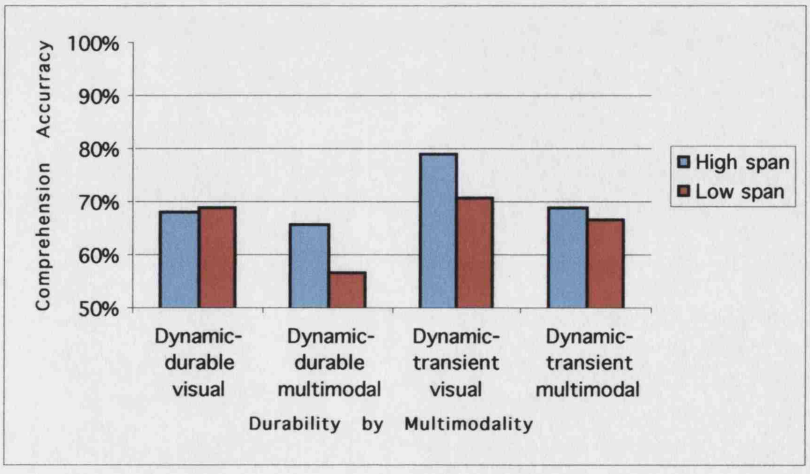
Table 6.2

Experiment 2a - Full Data. Mean Comprehension Rate (CR) for the Complex Conditions (%): By Span, Durability and Multimodality (Standard Errors in Parentheses)

WM Capacity	Durability by Multimodality			
	dynamic-durable visual	dynamic-durable multimodal	dynamic-transient visual	dynamic-transient multimodal
High	68%	66%	79%	69%
	(2%)	(4%)	(3%)	(4%)
Low	69%	57%	71%	67%
	(3%)	(5%)	(4%)	(5%)
Mean	69%	62%	75%	68%

Figure 6.2

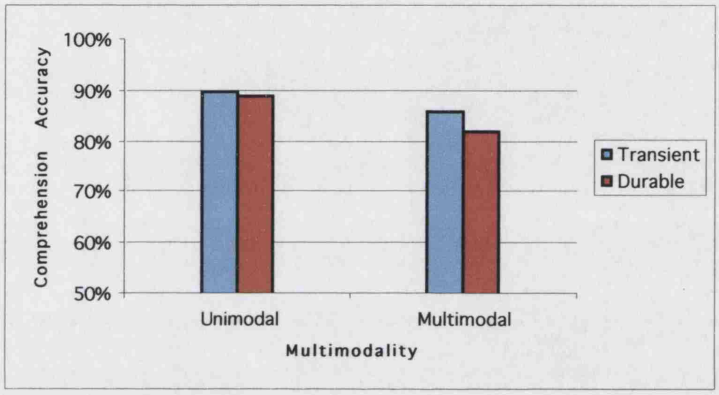
Experiment 2a - Full Data. Mean Comprehension Rate (CR) for the Complex Conditions (%): By Span, Durability and Multimodality



The investigation of this interaction involved a reduced analysis of variance at each level of the complexity variable. In contrast to the experimental predictions, for simple sentences, the main effect of multimodality reached significance: comprehension rates in the multimodal conditions (84%) were 5% lower than the rates found in the two visual presentation conditions (89%), ($F(1, 14) = 9.943$; $p < .01$). On the other hand and as predicted by the MMUM, the 3% advantage of the transient conditions (88%) over the durable conditions (85%) failed to reach significance ($F(1, 14) = 1.947$). Figure 6.3 shows that for simple sentences, these variables do not interact ($F(1, 14) = .636$). Furthermore, the interaction between durability, multimodality and span failed to reach significance and was not investigated further ($F(1, 14) = 2.533$).

Figure 6.3

Experiment 2a - Full Data. Mean Comprehension Rate (CR) for the Simple Conditions (%): By Durability and Multimodality



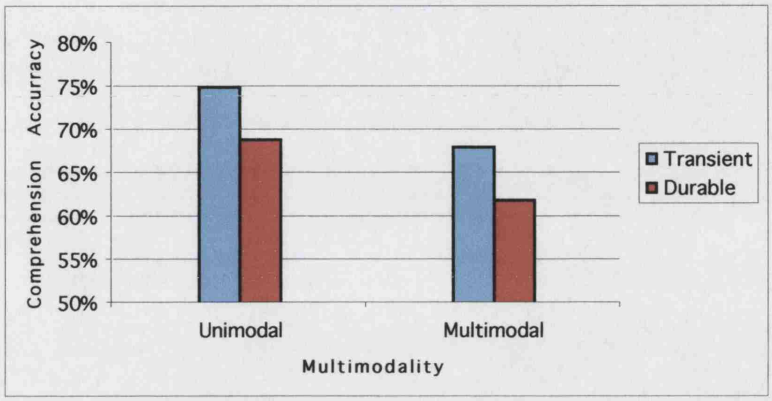
This pattern of results cannot be accounted for by the MMUM. The model predicted that a multimodal presentation of simple sentences would slightly improve performance if the visual and the auditory stimuli are presented together at the same rate. As suggested in Chapter 4, the simultaneous cross-modal activation of the lexical-access sub-system in the dynamic-transient multimodal condition was expected to feed the a-modal sub-systems with consistent representations of the verified words, resulting in a small reduction of user cost. In the multimodal dynamic-durable condition, the spoken continuation of the sentence was expected to slightly interfere with the collection of early sentential components by means of regressive eye-movements. However, according to the model, synchronous processing was assumed to be easily restored though refocusing attention on the "leading edge" of the visual display. In any case, the absence of a significant interaction between durability and multimodality for the simple conditions suggests that the multimodal interference effect took place for both durability conditions.

For the complex sentences, the interaction between durability, multimodality and span also failed to reach significance and was not investigated further ($F(1, 14) = 2.247$). However, the main effect of multimodality was nearly significant: comprehension rates in the multimodal conditions (65%) were 7% lower than the rates found for the two visual presentation conditions (72%), ($F(1, 14) = 4.403$; $p < .06$). Of greater interest, comprehension rates in the durable conditions (65%) were 7% lower than

the rates found for the transient conditions (72%), yielding a significant cost of durability for complex sentences ($F(1, 14) = 7.067$; $p < .03$). Figure 6.4 shows that contrary to the model's predictions, for complex sentences both effects are additive and fail to interact ($F(1, 14) = .129$).

Figure 6.4

Experiment 2a - Full Data. Mean Comprehension Rate (CR) for the Complex Conditions (%): By Durability and Multimodality



The absence of a significant interaction between durability and multimodality and specifically, the inferior performance of the dynamic-durable visual condition over the dynamic-transient visual condition that was identified for complex sentences cannot be accounted for by the MMUM. According to the model, regressive eye-movements were not expected to assist the interpretation of complex sentences in the dynamic-durable visual condition; subjects were assumed to have insufficient resources to recollect previously processed information without losing later sentential parts. Therefore, their comprehension rates were expected to equal those of the dynamic-transient visual condition: for both dynamic forms of presentation, the syntactic parser was not expected to have sufficient resources available for the delayed assignment of thematic roles required for the comprehension of complex sentences.

Response times

The time it took to comprehend each sentence was collected for each subject. Response times values were analysed in a repeated measure analysis of variance[41]. Complexity and span formed the

[41] An exploration of the data revealed that for the sub-conditions created by durability, multimodality and complexity, the assumption of normality was not met for four sub-conditions. The assumption of normality was not met for simple sentences in the dynamic-durable visual condition (Shapiro-Wilk (16) = .634; $p < .02$), for simple sentences in the dynamic-durable multimodal condition (Shapiro-Wilk (16) = .851; $p < .02$), for simple sentences in the dynamic-transient visual condition (Shapiro-Wilk (16) = .856; $p < .02$) and for complex sentences in the dynamic-transient visual condition (Shapiro-Wilk (16) = .867; $p < .03$). Furthermore, the assumption of homogeneity of variance was not met for the dynamic-durable multimodal conditions. For the sub-conditions created by durability, multimodality and span, the assumption of normality was not met for the low span subjects in the dynamic-durable visual condition (Shapiro-Wilk (14) = .842; $p < .02$), and for high

between-subjects variables while durability and multimodality formed the within-subjects variables. Response times for simple sentences (1632 ms) were significantly faster than those obtained for complex sentences (2047ms) ($F(1, 28) = 10.981$; $p < .01$). No other effect reached significance. The effect of complexity cannot shed light on the pattern of results found for the comprehension rate measure.

The role of the Line variable in processing sentences presented in a dynamic-durable form

The significant speech interference effect found for simple sentences and the absence of a significant interaction between durability and multimodality for complex sentences conflict with the predictions of the MMUM. It was decided to investigate other potential variables that might have played a role in this experiment, before assuming that the MMUM was incorrect and should be revised.

A prime candidate for such a variable was the number of lines it took to present sentences in the dynamic-durable conditions. Using a 15 inch monitor, more than half of the sentences were presented over two lines. If, towards the end of the first line, subjects could not predict whether a short word would appear at the end of the same line, or a longer word at the beginning of the next line, this might partially explain the results reported earlier. For simple sentences, this might explain the speech interference effect identified in the durable conditions: when subjects misjudged the location in which the next word was to appear, the normal continuation of the spoken sentence might have impaired their ability to refocus on the "leading edge" of the visual sentence. For complex sentences, this might explain the advantage of a dynamic-transient visual presentation over a dynamic-durable one: although the dynamic-transient format does not enable subjects to recollect previous information using regressive eye-movements, the location of each presented word is certain and so, does not distract from the highly demanding sentence processing task.

Item analysis

Comprehension rates were collected for each sentence. Complexity formed the within-items variable while line (three levels: transient; N=40, single-line; N=16, double-line; N=24) and multimodality formed the between-items variables. The pattern of the collected data did not warrant the use of parametric tests[42]. However, non-parametric tests cannot control for the effect of specific variables on

---

span subjects in the dynamic-transient visual condition (Shapiro-Wilk (18) = .795; $p < .01$). An additional exploration used the square root values of the response times data. This transformation improved the normality values (the assumption of normality was not met for only two sub-conditions created by durability, multimodality and complexity and for one sub-condition created by durability, multimodality and span). Meeting the assumption of homogeneity of variance also improved considerably and was now met for all presentation conditions created by the complexity, dynamism and multimodality variables and for three presentation conditions created by the span, dynamism and multimodality variables. It was decided to conduct the analyses of variance over the transformed values of the response times measure rather than the original values. Note that squared values were used to transform the results to the original units.

[42] An exploration of the items' data revealed that for the sub-conditions created by complexity and line, the assumption of normality was not met for complex sentences in the transient conditions (Shapiro-Wilk (40) =

the dependent measure. One such variable was the pragmatic complexity value of an item (see Chapter 2). Its influence on the comprehension of the sentence materials had to be assessed; a procedure that cannot be performed in the subjects' analysis. Using parametric tests, the effect of the pragmatic complexity on the comprehension rate measure could be established (by means of regression). Moreover, the effect of the pragmatic complexity on the comprehension rate measure could be controlled for by holding it constant (using the variable as a covariate). Thus, in spite of the fact that the assumptions of normality and of homogeneity of variance were not met as required, it was decided to perform the item analysis using parametric tests.

An analysis of variance was thus conducted, using the pragmatic complexity variable as a covariate. The analysis yielded the following effects: a significant effect of pragmatic complexity ($F(1, 73) = 7.779$; $p < .01$) and a significant interaction between pragmatic complexity and (syntactic) complexity ($F(1, 73) = 4.455$; $p < .05$). The source of this interaction was investigated separately by means of linear regressions. A significant negative relationship was found between pragmatic complexity and comprehension rate for both complex and simple sentences: the higher the pragmatic complexity value of the sentence, the lower the comprehension. This effect was stronger for complex sentences than for simple sentences:

Predicted $CR_{complex} = .911 - .024 \times$ pragmatic complexity + error
$(F(1, 79) = 9.096; p < .01)$
Predicted $CR_{simple} = .973 - .011 \times$ pragmatic complexity + error
$(F(1, 79) = 4.668; p < .05)$

Holding the pragmatic complexity variable constant, the effect of line approached significance in the item analysis ($F(2, 73) = 2.430$; $p < .1$). The data indicates that whereas comprehension rates of sentences presented in the single-line conditions (82%) did not differ from those found in the transient conditions (80%), comprehension rates of sentences presented over two lines were lower (72%). A partial contrast was conducted to assess the superiority of the single-line sentences over the double-line sentences using the LSD test. The differences approached significance ($p < .07$). It was found necessary to continue the investigation of the experimental results with separate subjects' analyses; one for the single-line sentences and the other for the double-line sentences.

.908; $p < .01$) and for simple sentences in all line conditions (transient: Shapiro-Wilk (40) = .798; $p < .01$, single-line: Shapiro-Wilk (16) = .628; $p < .01$, double-line: Shapiro-Wilk (24) = .824; $p < .01$).
For the sub-conditions created by complexity and multimodality, the assumption of normality was not met for complex sentences in the multimodal conditions (Shapiro-Wilk (40) = .896; $p < .01$) and for simple sentences in both the unimodal conditions (Shapiro-Wilk (40) = .766; $p < .01$) and the multimodal conditions (Shapiro-Wilk (40) = .848; $p < .01$). Furthermore, the assumption of homogeneity of variance was not met for complex sentences.

Subjects analyses: single-line

Comprehension rates

Due to the effect of the pragmatic complexity variable on comprehension rates, six sentences were selected for each presentation condition, with a similar mean of pragmatic complexity for each condition. Comprehension rates were collected for each subject and were analysed using non-parametric tests[43] (see Appendix C). The span variable was excluded from the analysis due to the small number of sentences per subject[44].

For the reduced single-line data set, comprehension rates of simple sentences (91%) were higher than those found for complex sentences (70%) (Mann-Whitney U = 18.50; Wilcoxon W = 154.50; Z = -4.152; $p < .01$). Also, the 7% advantage of the unimodal conditions (84%) over the multimodal conditions (77%) yielded a significant effect of multimodality (Z = -2.488; $p < .02$). Contrary to the full data set, the analysis did not find a significant effect of durability (81% for the durable conditions vs. 80% for the transient conditions) (Z = -.660). In addition, the interaction between the durability and multimodality reached significance (Z = -3.512; $p < .01$) (see Table 6.3).

Table 6.3

Experiment 2a – Single-Line. Mean Comprehension Rate (CR) for all Conditions (%): By Durability and Multimodality (Standard Errors in Parentheses)

	Multimodality		
Durability	Unimodal	Multimodal	Mean difference
Durable	79%	83%	-4%
	(3%)	(2%)	
Transient	89%	71%	18%
	(2%)	(4%)	
Mean difference	-10%	12%	

The source of the effect is the 18% redundancy cost found for the transient conditions (Z = -3.622; $p < .01$), rather than the small *gain* found for the durable conditions (Z = -.823).

Significantly for the purpose of research, the interaction between durability, multimodality and complexity reached significance. Table 6.4 presents the mean comprehension rate created by the three variables. The relationships between these means can be seen more readily in Figures 6.5 and 6.6.

[43] An exploration of the data revealed that for the sub-conditions created by the durability, multimodality and complexity, the assumption of normality was not met for seven out of eight sub-conditions. Furthermore, the assumption of homogeneity of variance was not met for the dynamic-transient multimodal condition.

[44] Note that due to the small number of sentences per subject, comprehension rates of complex sentences in the dynamic-transient multimodal condition did not depart from a chance level of performance ($t_{15} = .921$).

Table 6.4

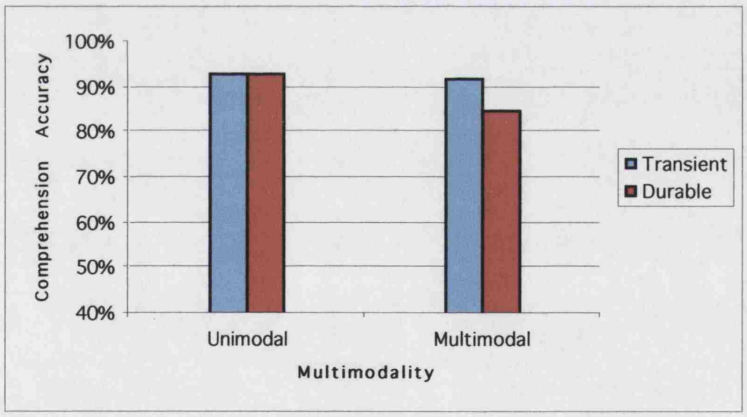
Experiment 2a – Single-Line. Mean Comprehension Rate (CR) for the Complexity Conditions (%): By Durability and Multimodality (Standard Errors in Parentheses)

Complexity	Durability by Multimodality			
	dynamic-durable visual	dynamic-durable multimodal	dynamic-transient visual	dynamic-transient multimodal
Simple	93%	92%	93%	85%
	(4%)	(3%)	(3%)	(5%)
Complex	66%	75%	84%	56%
	(4%)	(3%)	(3%)	(5%)
Mean difference	27%	17%	9%	29%

The investigation of this triple interaction involved a reduced analysis at each level of the complexity variable. For the simple conditions, neither an effect of durability nor an effect of multimodality were identified ($Z_{Durability} = -.998$), ($Z_{Multimodality} = -1.270$). Furthermore, the interaction between the two factors did not reach significance ($Z = -1.078$).

Figure 6.5

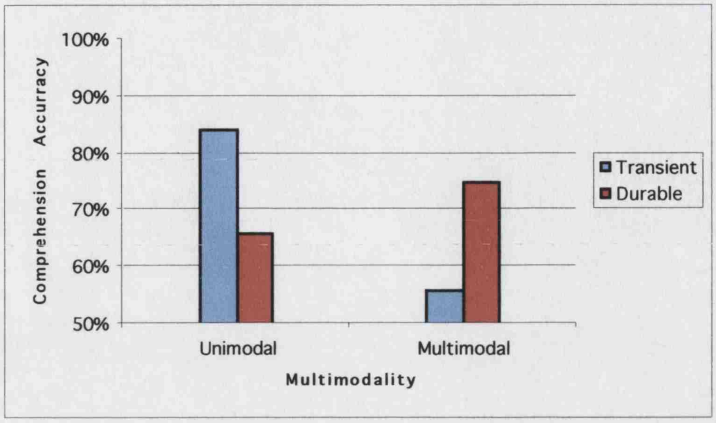
Experiment 2a – Single-Line. Mean Comprehension Rate (CR) for the Simple Conditions (%): By Durability and Multimodality



For the complex conditions, no effect of durability was identified ($Z = -.000$). However, the effect of multimodality reached significance ($Z = -2.153$; $p < .04$). Furthermore, the interaction between the two factors reached significance ($Z = -3.238$; $p < .01$). Simple main effects revealed that the source of the effect is in the 28% redundancy cost found in the transient conditions (56% in the dynamic-transient multimodal condition vs. 84% in the dynamic-transient visual condition) ($Z = -3.093$; $p < .01$). The 9% redundancy gain found in the durable conditions (75% in the dynamic-durable multimodal condition vs. 66% in the dynamic-durable visual condition) did not reach significance ($Z = -1.517$).

Figure 6.6

Experiment 2a –Single-Line. Mean Comprehension Rate (CR) for the Complex Conditions (%): By Durability and Multimodality



Response times

The time it took to comprehend each sentence for the reduced single-line data set was collected for each subject. These values were analysed using non-parametric tests[45]. The span variable was again excluded from the analysis. For the reduced single-line data set, response times for simple sentences (1575 ms) were significantly faster than those obtained for complex sentences (2073ms) (Mann-Whitney U = 35.00; Wilcoxon W = 171.00; Z = –3.360; $p < .01$). No other effect reached significance.

Interim summary

For simple sentences, the overall pattern of results is compatible with the predictions made by the MMUM. For complex sentences, the pattern of results differs substantially from the predicted effects. The model did not predict a speech interference effect for the dynamic-transient visual presentation of complex sentences, an effect clearly evident in this experiment. On the other hand, it predicted a speech-interference effect for complex sentences displayed by means of dynamic-durable visual presentation. Specifically, the incompatibility of the spoken information with the recollected visual words was expected to produce an interference effect at all levels of processing. Although the opposite result identified for complex sentences in the dynamic-durable conditions did not reach significance, the 9% redundancy gain may partially support the durability account suggested for

[45] An exploration of the data revealed that for the sub-conditions created by durability, multimodality and complexity, the assumption of normality was not met for simple sentences in the dynamic-durable visual condition (Shapiro-Wilk (16) = .764; $p < .01$), for simple sentences in the dynamic-durable multimodal condition (Shapiro-Wilk (16) = .839; $p < .05$) and for complex sentences in the dynamic-transient visual condition (Shapiro-Wilk (16) = .775; $p < .01$). Furthermore, for the sub-conditions created by durability, multimodality and span, the assumption of normality was not met for high span subjects in the dynamic-durable visual conditions (Shapiro-Wilk (17) = .883; $p < .05$), in the dynamic-durable multimodal conditions (Shapiro-Wilk (17) = .861; $p < .05$) and in the dynamic-transient visual condition (Shapiro-Wilk (17) = .818; $p < .01$).

experiment 1. This account suggests that users are able to fully utilise the affordance of a durable visual presentation to recollect early sentential components by switching attention between modalities so as to assign thematic roles across them. However, the significant redundancy cost identified for the dynamic-transient complex sentences suggests that this interpretation is premature. If multimodality in itself impairs the processing of complex sentences, can it facilitate comprehension of complex sentences presented by means of durable text? Alternatively, can this pattern of result simply imply that one cannot reliably interpret a set of results based on 6 sentences per subject? A replication of this experiment that avoids the confounding line variable is necessary to decide between these two options.

Subjects analyses: double-line

Comprehension rates

Ten sentences were selected for each presentation condition, with a similar mean of pragmatic complexity for each condition. Comprehension rates were collected for each subject and were analysed using non-parametric tests[46]. The span variable was excluded from the analysis due to the small number of sentences per subject[47]. For the reduced double-line data set, a significant triple interaction between durability, multimodality and complexity was not found (Mann-Whitney U = 93.50; Wilcoxon W = 229.50; Z = -1.317). Table 6.5 presents the mean comprehension rate created by the three variables. The relationships between these means can be seen more readily in Figures 6.7 and 6.8.

Table 6.5

Experiment 2a – Double-Line. Mean Comprehension Rate (CR) for the Complexity Conditions (%): By Durability and Multimodality (Standard Errors in Parentheses)

Complexity	Durability by Multimodality			
	dynamic-durable visual	dynamic-durable multimodal	dynamic-transient visual	dynamic-transient multimodal
Simple	89%	75%	88%	88%
	(3%)	(4%)	(3%)	(4%)
Complex	68%	48%	81%	68%
	(3%)	(4%)	(3%)	(4%)
Mean difference	21%	27%	7%	20%

[46] An exploration of the data revealed that for the sub-conditions created by durability, multimodality and complexity, the assumption of normality was not met for simple sentences in the dynamic-durable visual condition (Shapiro-Wilk (16) = .840; $p < .01$), in the dynamic-transient visual condition (Shapiro-Wilk (16) = .878; $p < .05$) and in the dynamic-transient multimodal condition (Shapiro-Wilk (16) = .840; $p < .01$) = .876; $p < .04$). Furthermore, the assumption of homogeneity of variance was not met for the dynamic-transient multimodal condition.

[47] Note that due to the small number of sentences per subject, comprehension rates of complex sentences in the dynamic-durable multimodal condition did not depart from a chance level of performance ($t_{15} = -.409$).

Figure 6.7

Experiment 2a – Double-Line. Mean Comprehension Rate (CR) for the Simple Conditions (%): By Durability and Multimodality

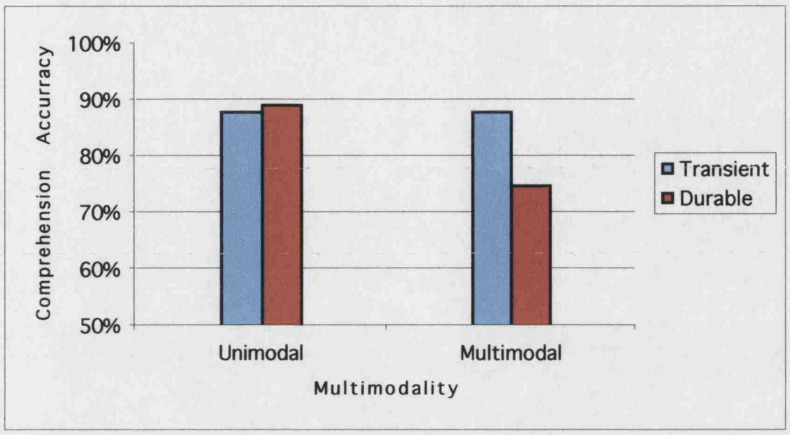
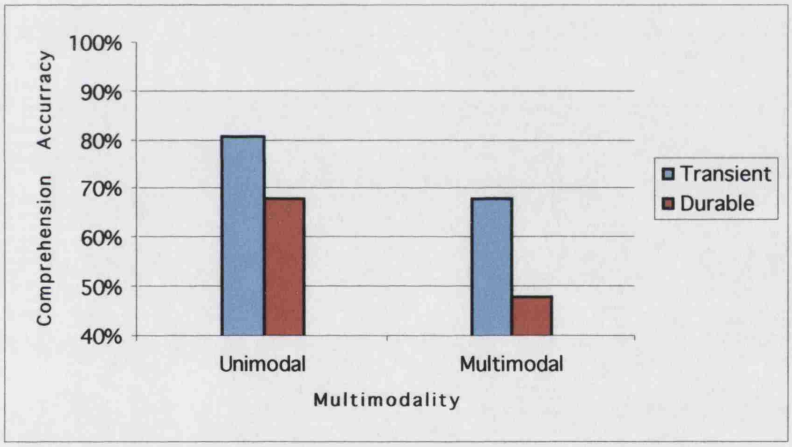


Figure 6.8

Experiment 2a – Double-Line. Mean Comprehension Rate (CR) for the Complex Conditions (%): By Durability and Multimodality



Overall, comprehension rates of simple sentences (85%) were higher than those of complex sentences (66%) (Mann-Whitney U = 24.50; Wilcoxon W = 160.50; Z = -3.913; $p < .01$), yielding a significant effect of complexity. Similar to the results reported for the single-line, the 12% advantage of the unimodal conditions (82%) over the multimodal conditions (70%) yielded a significant effect of multimodality (Z = -3.693; $p < .01$). However, in this analysis, the interaction between multimodality and complexity reached significance (Mann-Whitney U = 71.50; Wilcoxon W = 207.50; Z = -2.152; $p < .04$). Simple main effects revealed that this interaction was not reliable. For complex sentences, the advantage of the unimodal conditions (75%) over the multimodal conditions (58%) was highly significant (Z = -2.848; $p < .01$). For simple sentences, the smaller advantage of the unimodal conditions (88%) over the multimodal conditions (81%) also reached significance (Z = -2.346; $p < .03$).

Furthermore, in contrast to the single-line results, the 11% advantage of the transient conditions (81%) over the durable conditions (70%) yielded a significant effect of durability ($Z = -3.913$; $p < .01$). The analysis also yielded a significant interaction between durability and complexity (Mann-Whitney U = 62.00; Wilcoxon W = 198.00; $Z = -2.508$; $p < .02$). The large advantage of the transient conditions (74%) over the durable conditions (58%) was highly significant for complex sentences ($Z = -3.423$; $p < .01$). For the simple conditions, the smaller advantage of the transient conditions (88%) over the durable conditions (82%) approached significance ($Z = -1.940$; $p < .06$).

Finally, the analysis yielded a significant interaction between durability and multimodality ($Z = -2.848$; $p < .01$). For the durable conditions, the large advantage of the unimodal conditions (79%) over the multimodal conditions (62%) was highly significant ($Z = -4.164$; $p < 01$). For the transient conditions, the smaller advantage of the unimodal conditions (84%) over the multimodal conditions (78%) approached significance (-1.891; $p < .07$).

Response times

The time it took subjects to comprehend each sentence for the reduced double-line data set was collected for each subject. These values were analysed using non-parametric tests[48]. The span variable was again excluded from the analysis.

For the reduced double-line data set, response times for simple sentences (1639 ms) were again significantly faster than those obtained for complex sentences (2055 ms) (Mann-Whitney U = 51.00; Wilcoxon W = 187.00; $Z = -2.902$; $p < .01$). Also, the triple interaction between durability, multimodality and complexity approached significance (Mann-Whitney U = 76.00; Wilcoxon W = 212.00; $Z = -1.960$; $p < .06$). The investigation of this triple interaction involved a reduced analysis of variance at each level of the complexity variable. The interaction between durability and multimodality reached significance only for complex sentences ($Z = -2.017$; $p < .05$). Simple main effects reveal that the response times in the dynamic-transient multimodal condition (1845 ms) were faster than those found for the dynamic-transient visual condition (2138 ms). This difference approached significance ($Z = -1.706$; $p < .1$), raising the possibility that had the response times in the dynamic-transient complex conditions been equal, speech interference might have been smaller. In contrast, the response times in the dynamic-durable visual condition (2018 ms) did not differ significantly from the response times in the dynamic-durable multimodal condition (2218 ms) ($Z = -1.086$).

---

[48] An exploration of the data revealed that for the sub-conditions created by durability, multimodality and complexity, the assumption of normality was not met for three sub-conditions: for simple sentences in the dynamic-durable visual condition (Shapiro-Wilk (16) = .626; $p < .01$), for simple sentences in the dynamic-transient visual condition (Shapiro-Wilk (16) = .848; $p < .05$) and for complex sentences in the dynamic-transient visual condition (Shapiro-Wilk (16) = .775; $p < .01$). Also, the assumption of homogeneity of variance was not met for the dynamic-durable multimodal condition.

Interim summary:

The results of various analyses imply that for the full data set, the limitations of a 15 inch monitor impaired subjects' ability to predict the location in which words were to appear in the dynamic-durable conditions since more than half of the sentences were presented over two lines. Towards the end of the first line, subjects could not predict whether a short word would appear at the end of the same line, or a longer word at the beginning of the next line. It is suggested that this lack of predictability is responsible for the unaccountable results of the dynamic-durable conditions in the full data set.

Specifically, the significant interaction between the durability and multimodality variables found in the double-line analysis suggests that, whereas for the transient conditions concurrent speech impairs performance, the magnitude of interference is significantly higher when the location of the visual words cannot be predicted in advance. Furthermore, although each of these variables interacted with the complexity variable, the absence of a triple interaction between durability, multimodality and complexity suggests that the processing of simple and complex sentences did not differ significantly in this respect. Figures 6.7 and 6.8 clearly demonstrate that whereas the source of the speech interference effect for the dynamic-transient conditions is the large difference found for complex sentences, for the dynamic-durable conditions the effect was identified for both simple and complex sentences.

For the single-line data set, the pattern of results differed substantially. For simple sentences, neither durability nor multimodality affected comprehension rates. For complex sentences, a speech interference effect was identified only in the dynamic-transient conditions; processing complex sentences displayed by means of a dynamic-durable text was not affected by the concurrent speech. Due to the small number of single-line sentences in this experiment, this result was found unreliable. It was decided to run the experiment again using a 17 inch monitor.

6.2 Experiment 2b

In the replication of experiment 2a, two computers were used: the transient conditions were run on a PowerPC 5200/75 for which the built-in 15 inch Apple display was sufficient, and the durable conditions were run on a PowerPC G3/300 for which a 17 inch Compaq display was used. All sentences in the durable conditions were now presented in one line. For the experimental hypotheses and the full method, see sections 6.1.2 and 6.1.3. The experimental results for 32 different subjects (16 high span and 16 low span) are reported next.

6.2.1 Results & Discussion

Subjects' Analyses

Comprehension rates

Comprehension rates were collected for each subject and were analysed in a repeated-measure analysis of variance. Complexity and span formed the between-subjects variables while durability and multimodality formed the within-subjects variables[49]. Comprehension rates of simple sentences (85%) were 17% higher than those found for complex sentences (68%) yielding a main effect of complexity ($F(1, 28) = 28.511$; $p < .01$)[50]. Moreover, the analysis produced a significant interaction between complexity, durability and multimodality ($F(1, 28) = 4.465$; $p < .05$). Tables 6.6 and 6.7 present the mean comprehension rates created by the durability and the multimodality variables for each complexity level. The relationship between these means can be seen more readily in Figures 6.9 and 6.10.

Table 6.6

Experiment 2b. Mean Comprehension Rate (CR) for Simple Sentences (%): By Durability and Multimodality (Standard Errors in Parentheses)

	Multimodality		
Durability	Unimodal	Multimodal	Mean difference
Durable	84%	83%	1%
	(3%)	(2%)	
Transient	88%	85%	3%
	(3%)	(3%)	
Mean difference	-4%	-2%	

[49] An exploration of the comprehension rate data revealed that for the sub-conditions created by complexity, durability and multimodality, the assumption of normality was met for all but the dynamic-transient multimodal complex condition (Shapiro-Wilk (16) = .861; $p < .03$). Furthermore, for the sub-conditions created by span, durability and multimodality, the assumption of normality was not met for low span subjects in the dynamic-durable visual condition (Shapiro-Wilk (16) = .860; $p < .03$) and the dynamic-transient visual condition (Shapiro-Wilk (16) = .768; $p < .01$). The assumption of homogeneity of variance was met for all sub-conditions.

[50] Averaged across span, comprehension rates of complex sentences were higher than chance level in all presentation conditions ($t_{15} = 3.962 - 9.271$, $p < .05$).

Figure 6.9

Experiment 2b. Mean Comprehension Rate (CR) for the Simple Conditions (%): By Durability and Multimodality

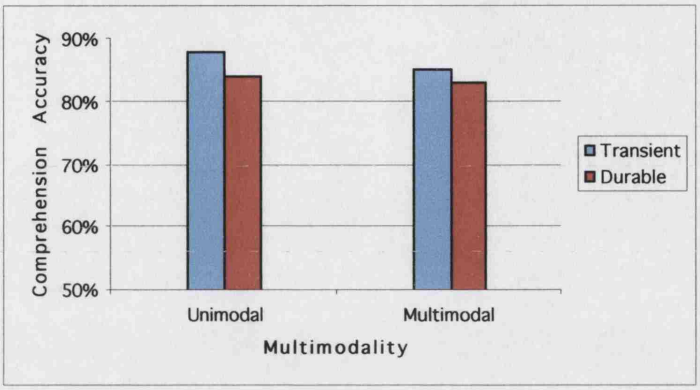


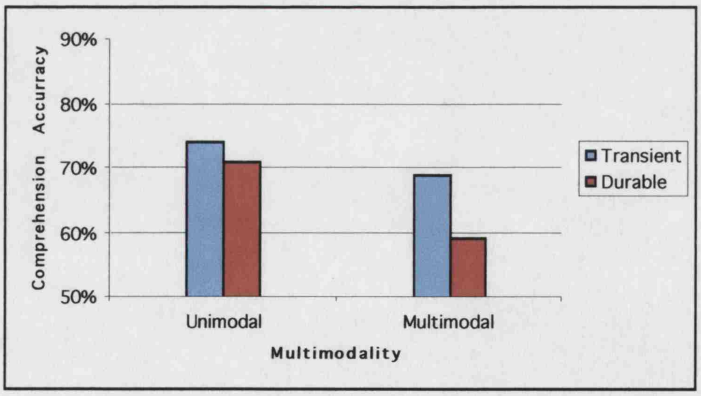
Table 6.7

Experiment 2b. Mean Comprehension Rate (CR) for Complex Sentences (%): By Durability and Multimodality (Standard Errors in Parentheses)

	Multimodality		
Durability	Unimodal	Multimodal	Mean difference
Durable	71%	59%	12%
	(3%)	(2%)	
Transient	74%	69%	5%
	(3%)	(3%)	
Mean difference	-3%	-10%	

Figure 6.10

Experiment 2b. Mean Comprehension Rate (CR) for the Complex Conditions (%): By Durability and Multimodality



The investigation of this interaction involved a reduced analysis of variance at each level of the complexity variable. For the simple conditions, results supported the experimental predictions: neither an effect of durability nor an effect of multimodality were identified in the analysis ($F_{Durability}$

(1, 14) = 2.637), ($F_{Multimodality}$ (1, 14) = .535). Furthermore, the interaction between these two factors did not reach significance (F (1, 14) = .529).

For the complex conditions, the pattern of results is only partially compatible with the experimental predictions. The 6% advantage of the transient conditions (71%) over the durable conditions (65%) yielded a significant effect of durability (F (1, 14) = 8.511; $p < .02$). In addition, the 8% advantage of the unimodal conditions (72%) over the multimodal conditions (64%) yielded a significant effect of multimodality (F (1, 14) = 20.852; $p < .01$). Finally, the interaction between durability and multimodality reached significance (F (1, 14) = 4.628; $p < .05$) (see Table 6.7 and Figure 6.10).

Simple main effects revealed that the source of the effect is the higher redundancy cost found in the durable conditions (F (1, 15) = 21.593; $p < .01$) relative to the cost found in the transient conditions (F (1, 15) = 4.615; $p < .05$). For the durable conditions, this result supports the predictions made by the MMUM; this clear-cut evidence of speech interference contrasts with the (insignificant) facilitation observed in the single-line analysis of experiment 2a. It seems that for durable presentation of complex sentences, the asynchrony of the spoken information, that occurs with visual regressions following normal processing breakdown, produces an interference effect at all levels of processing for all users. Since the SAS does not have sufficient resources available to supervise the competition between the language sub-systems for activation and the coordination of information between modalities, the addition of speech impairs comprehension of complex sentences. For the transient condition, the speech interference effect replicates the trends found in all three analyses of experiment 2a. The compelling evidence that multimodality in itself impairs comprehension of complex sentences necessitates a fundamental modification of the MMUM. Not only that a simultaneous cross-modal activation of the lexical-access sub-system is negligible when processing complex sentences, but also that the delayed assignment of thematic roles (necessary for the comprehension of complex sentences) is impaired by the redundant speech. This implies that for complex sentences, speech interferes not only when regressive eye-movements are present. Mental recollection of verbal information is also impaired by concurrent speech, although the magnitude of the interference is reduced relative to a physical recollection of visual information [51].

---

[51] An additional analysis was conducted at the request of the examiners of this dissertation. This analysis included the order in which subjects performed the four presentation conditions as an additional between-subjects variable (version number) to make sure that the experimental results were unaffected by order effects (see Method, section 6.1.3). Whereas most of the results of this experiment were unaffected by the order in which subjects performed the four presentation conditions, the interaction between complexity, durability and multimodality only approached significance in this analysis (F (1, 16) = 3.069; $p < .1$). Specific effects differed for both complexity conditions. For the simple sentences, both the effect of multimodality and the interaction between durability and multimodality remained insignificant ($F_{multimodality}$ (1, 8) = .781), ($F_{durability \times multimodality}$ (1, 8) = .750). In contrast, the 3% advantage of the transient conditions (86.5%) over the durable conditions (83.5%) yielded an effect of durability (F (1, 8) = 11.571, $p < .01$). For the complex sentences, the effect of durability and the effect of multimodality remained significant ($F_{durability}$ (1, 8) = 8.456, $p < .05$), ($F_{multimodality}$ (1, 8) = 33.573, $p < .01$), but the interaction between the two factors failed to reach significance (F (1, 8) = 2.307). The main difference between the absence of a significant interaction between durability and multimodality in

On average, the pattern of this triple interaction yielded a significant interaction between multimodality and complexity ($F(1, 28) = 5.750$; $p < .05$). The mean comprehension rates created by these variables are shown in Table 6.8 and Figure 6.11. Simple main effects revealed that the source of the effect is the redundancy cost found for complex sentences ($F(1, 14) = 20.852$; $p < .01$). For simple sentences, no evidence of an effect of multimodality was identified ($F(1, 14) = .535$).

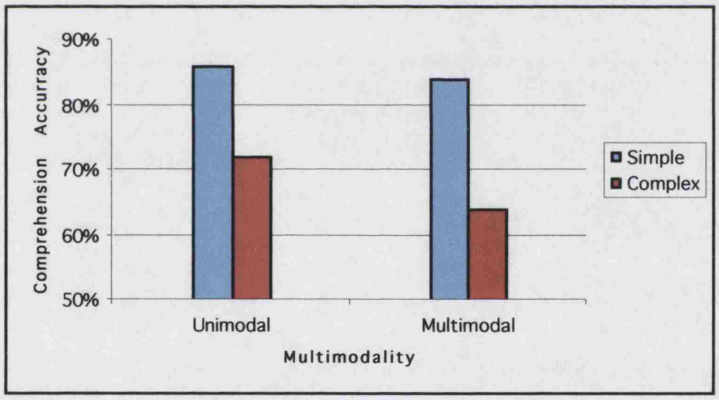
Table 6.8

Experiment 2b. Mean Comprehension Rate (CR) for Complexity Conditions (%): By Multimodality (Standard Errors in Parentheses)

	Multimodality		
Complexity	Unimodal	Multimodal	Mean difference
Simple	86%	84%	2%
	(3%)	(2%)	
Complex	72%	64%	8%
	(3%)	(2%)	
Mean difference	14%	20%	

Figure 6.11

Experiment 2b. Mean Comprehension Rate (CR) for Complexity Conditions (%): By Multimodality



the version-based analysis and its presence in the original analysis is the implication of the version-based results that speech interferes with the comprehension of complex sentences regardless of the durability of a dynamic visual text. This difference does not have serious implications for the MMUM. Similar to the original results, these results support the predictions made by the MMUM for the durable conditions and require a modification of the predictions concerning dynamic-transient multimodality.

Note that the number of subjects in each version (between 1 to 3 subjects) was too small to be confident that the alternative results are more correct than the original ones. For this number of subjects, the Box's M test of equality of covariance matrices could not be computed. Furthermore, the assumption of homogeneity of variance was not met for all conditions. Consequently, results of the version-based analysis do not seem highly reliable. However, some of the results concerning the verbal WM capacity variable suggest that practice effects might have operated in this experiment and, as such, require further investigation of the version-based analysis. These will be pinpointed later in this section.

Finally, the analysis of variance produced over the complexity conditions yielded a significant effect of durability ($F(1, 28) = 10.774$; $p < .01$) and a significant effect of multimodality ($F(1, 28) = 12.343$; $p < .01$). The sources of these effects are the large differences obtained in the complex conditions.

Individual differences in verbal WM

Consistent with the experimental predictions, comprehension rates of high span subjects (82%) were 11% higher than those found for low span subjects (71%), yielding a main effect of span ($F(1, 28) = 11.453$; $p < .01$). The expected relationships between the span variable and the other variables were not confirmed, however, by the experimental results[52].

---

[52] As noted earlier, the results of the version-based analysis do not seem highly reliable. However, this analysis yielded some significant effects relating to the verbal WM capacity variable, which suggest that some practice effects might have operated in this experiment. First, the analysis yielded a significant interaction between durability, span and version ($F(3, 16) = 3.590$; $p < .05$). For high span subjects, version number did not interact with the durability variable ($F(3, 8) = 1.781$). However, for low span subjects, the interaction between version and durability reached significance ($F(3, 8) = 10.412$; $p < .01$). The differences between the comprehension rates of low span subjects in the durable and the transient conditions are summarised next for each version:

Experiment 2b - Version-based Analysis
Mean Comprehension Rate (CR) for Low Span Subjects in all Order Versions (%): By Durability

	Durability		
Version number	Durable	Transient	Difference
1	74%	79%	-5%
2	61%	71%	-10%
3	72%	80%	-8%
4	70%	67%	+3%

This table shows an obscure asymmetrical transfer (the effect upon B being preceded by A is different from the effect upon A of being preceded by B) for low span subjects. For versions 1 and 3, in which durable conditions preceded transient conditions (see Method, section 6.1.3), comprehension in the durable conditions was lower than comprehension in the transient conditions. Both effects only approached significance ($F_{version\ 1}(1, 2) = 16.200$; $p < .06$), ($F_{version\ 3}(1, 2) = 56.333$; $p < .09$). For versions 2 and 4, in which transient conditions preceded durable conditions, comprehension in the transient conditions was either higher than the comprehension in the durable conditions (version 2) or *lower* than the comprehension in the durable conditions (version 4). Again, both effects only approached significance ($F_{version\ 2}(1, 2) = 9.846$; $p < .09$), ($F_{version\ 4}(1, 2) = 9.600$; $p < .06$). The small number of span subjects in each version might suggest that the approaching significance values are a matter of poor experimental power. Care should be taken in the interpretation of the effects concerning durability and span.

Specifically, the MMUM predicted that the span variable would interact with the durability and multimodality variables only for complex sentences. For simple sentences presented in a durable form, the model predicted a slightly larger facilitation of multimodality for high span subjects than for low span subjects. High span subjects were assumed to have a sufficient amount of resources to accommodate both storage and computational demands of simple sentences and to maintain successful divided attention between modalities. For low span subjects, it was believed that the coupled outputs would compensate for any local interference caused by incompatible spoken output with the recollected visual information. Moreover, for transient presentation of simple sentences, the model predicted a slight facilitation of multimodality for both low and high span subjects.

The results indicate a different pattern of performance: for simple sentences, the analysis yielded an approaching significance interaction between span and multimodality ($F(1, 14) = 3.618$; $p < .08$)[53].

Second, the analysis yielded an interaction between multimodality, span and version ($F(3, 16) = 3.494$; $p < .05$). Again, for high span subjects, version number did not interact with the multimodality variable ($F(3, 8) = 1.627$). For low span subjects, the interaction between multimodality and version approached significance ($F(3, 8) = 3.323$; $p < .08$). The differences between the comprehension rates of low span subjects in the unimodal and the multimodal conditions are summarised next for each version:

Experiment 2b - Version-based Analysis
Mean Comprehension Rate (CR) for Low Span Subjects in all Order Versions (%): By Multimodality

	Multimodality		
Version number	Unimodal	Multimodal	Difference
1	80%	73%	7%
2	68%	65%	3%
3	84%	68%	16%
4	72%	66%	6%

Simple main effects reveal an indication of an asymmetrical transfer for the multimodality variable. For versions 3 and 4, in which multimodal conditions preceded unimodal conditions, comprehension in the multimodal conditions was lower than comprehension in the unimodal conditions. For version 3, the effect of multimodality was significant ($F(1, 2) = 243.000$; $p < .05$) and for version 4, the effect approached significance ($F(1, 2) = 9.000$; $p < .06$). In contrast, when unimodal conditions preceded multimodal conditions (version 1 and 2), the effect of multimodality did not reach significance ($F_{version\ 1}(1, 2) = 1.862$), ($F_{version\ 2}(1, 2) = 8.000$). Again, the small number of span subjects in each version might suggest that the approaching significance interaction between multimodality and version is a result of low experimental power. Care should be taken in the interpretation of the effects concerning multimodality and span.

The last effect concerning order version is a 5-ways interaction between durability, multimodality, complexity, span and version which approached significance ($F(1, 16 = 3.249$; $p < .06$). Due to the small number of span subjects in each version, it was decided that this complex interaction is unlikely reliable. The interaction was not analysed any further.

[53] The significance value was higher in the version based analysis ($F(1, 8) = 5.281$; $p < .06$).

Table 6.9 provides the mean comprehension rates obtained for the two variables and Figure 6.12 shows them graphically.

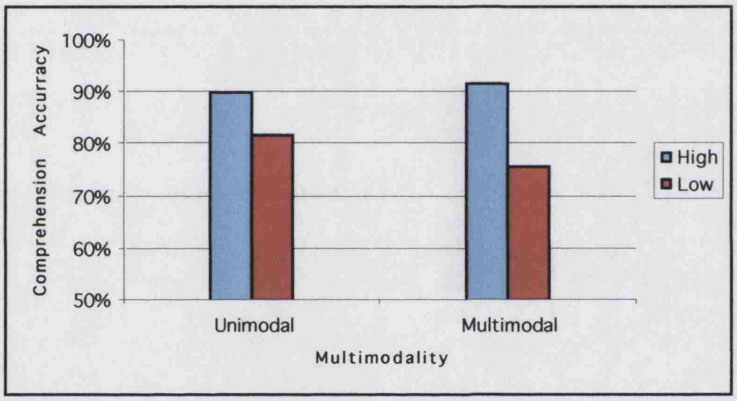
Table 6.9

Experiment 2b. Mean Comprehension Rate (CR) for Simple Sentences (%): By Span and Multimodality (Standard Errors in Parentheses)

	Multimodality		
Span	Unimodal	Multimodal	
High	90%	92%	-2%
	(4%)	(3%)	
Low	82%	76%	6%
	(4%)	(3%)	
	8%	16%	

Figure 6.12

Experiment 2b. Mean Comprehension Rate (CR) for Simple Sentences (%): By Span and Multimodality



Simple main effects show that this interaction is not reliable. Although concurrent speech impairs the comprehension of long-simple sentences for low span subjects by 6%, this interference effect is not significant (F (1, 7) = 2.910)[54]. For high span subjects, concurrent speech does not affect comprehension (F (1, 7) = .848). It is possible that having been exposed to a single structure in the simple condition, high span subjects hit the ceiling in their comprehension rates for the unimodal conditions. This might have defeated the attempt to reveal a slightly higher performance in the multimodal conditions for these subjects. The interpretation of the absence of speech interference effect for low span subjects is more difficult. On one hand, the absence of speech interference effect for low span subjects might simply confirm the predictions of the MMUM; that for simple sentences, coupled outputs compensate for any local interference caused by incompatible processes. Coupled

[54] In the version-based analysis, the effect approaches significance (F (1, 4) = 4.629; $p < .1$). However, the multimodality practice effect, identified for low span subjects across complexity conditions, supports the acceptance of the null hypothesis that the two groups of span come from populations with the same average mean in the simple condition.

outputs may compensate for incompatibility between speech output and the recollected visual information in a dynamic-durable multimodal presentation, and for incompatibility between "real-time" multimodal output and the "forgotten" intermediate computational products in a dynamic-transient multimodal presentation. However, this interpretation seems premature. An inspection of the comprehension rates found for simple sentences in experiment 2a shows that at least for the dynamic-transient conditions (for which results are assumed to be valid[55]) multimodality seems to impair comprehension for low span subjects (see Table 6.1 and Figure 6.1). Note that this observation is not based on a statistically significant effect. The same trend repeats in this experiment; Table 6.10 provides the mean comprehension rates obtained for the span, durability and multimodality variables for the simple sentences in experiment 2b.

Table 6.10

Experiment 2b. Mean Comprehension Rate (CR) for Simple Sentences (%): By Span, Durability and Multimodality (Standard Errors in Parentheses)

	Durability by Multimodality			
Span	dynamic-durable visual	dynamic-durable multimodal	dynamic-transient visual	dynamic-transient multimodal
High	88%	91%	91%	94%
	(4%)	(3%)	(4%)	(4%)
Low	79%	76%	84%	76%
	(4%)	(3%)	(4%)	(4%)
Mean difference	9%	15%	7%	18%

Looking at the results of both experiments raises suspicion that the absence of speech interference found for low span subjects in comprehending simple sentences may reflect a low experimental power; a result of the limited number of span subjects in each complexity condition. An additional analysis was therefore conducted to test the absence of speech interference in comprehending simple sentences for low span subjects using the data of the dynamic-transient conditions from both experiments 2a and 2b[56]. For the dynamic-transient simple conditions, the interaction between multimodality and span was significant ($F(1, 30) = 9.016$; $p < .01$). Simple main effects revealed that whereas comprehension of high span subjects was unaffected by the added speech ($F(1, 14) = .424$), the comprehension of low span subjects was significantly lower in the multimodal condition than in the unimodal condition ($F(1, 14) = 13.576$; $p < .01$).

[55] and for which, no order effects were identified

[56] It could be claimed that this analysis is problematic. Results of the dynamic-transient conditions in experiment 2a might have also been affected by the increased difficulty imposed by the double-line sentences in the dynamic-durable conditions. However, since no order effects were identified in experiment 2a, and since presentation was varied across uniform blocks, collecting data from both experiments may serve as a preliminary source of information regarding the validity of the results of experiment 2b.

This result requires a modification of the MMUM. It suggests that for low span subjects processing long-simple sentences in a dynamic-transient multimodal presentation, not only is a simultaneous cross-modal activation of the lexical-access sub-system negligible but that their SAS does not have sufficient resources to accommodate both storage and computational demands of the sentence and to maintain a successful coordination of information. The added speech seems to impair the mental recollection of intermediate representations by these subjects. More subjects are needed to verify a similar effect in the dynamic-durable conditions.

For complex sentences, the predicted interaction between span, durability and multimodality did not reach significance ($F(1, 14) = 2.361$)[57]. Table 6.11 presents the means created by the three variables. The relationships between these means can be seen in Figures 6.13 and 6.14. An inspection of the data shows that although statistically insignificant, the trend of the data partially supports this predicted interaction.

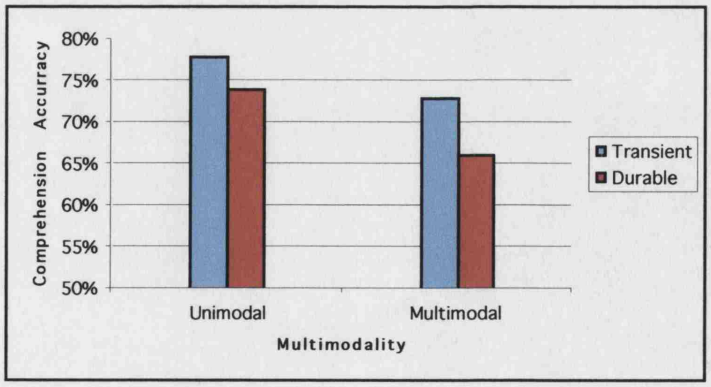
Table 6.11

Experiment 2b. Mean Comprehension Rate (CR) for Complex Sentences (%): By Span, Durability and Multimodality (Standard Errors in Parentheses)

Span	Durability by Multimodality			
	dynamic-durable visual	dynamic-durable multimodal	dynamic-transient visual	dynamic-transient multimodal
High	74% (4%)	66% (3%)	78% (4%)	73% (4%)
Low	68% (4%)	53% (3%)	69% (4%)	65% (4%)
Mean difference	6%	13%	9%	8%

Figure 6.13

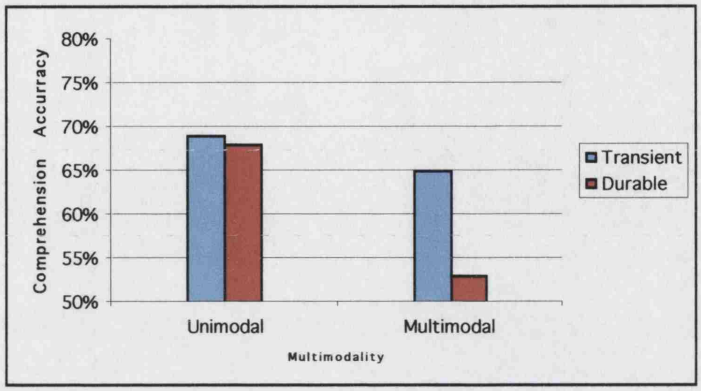
Experiment 2b. Mean Comprehension Rate (CR) of Complex Sentences for High Span Users (%): By Durability and Multimodality



[57] In the version based analysis, the effect approaches significance ($F(1, 8) = 3.505$; $p < .1$).

Figure 6.14

Experiment 2b. Mean Comprehension Rate (CR) of Complex Sentences for Low Span Users (%): By Durability and Multimodality



Further analyses revealed that for high span subjects, the only effect that approached significance was the multimodal interference effect (F (1,7) = 4.391; $p < .08$), (see Figure, 6.13). For low span subjects, the effects of durability, the effect of multimodality and the interaction between the two variables reached significance ($F_{durability}$ (1, 4) = 9.308; $p < .05$), ($F_{multimodality}$ (1, 4) = 29.867; $p < .01$), ($F_{durability \times multimodality}$ (1, 4) = 9.610; $p < .05$), (see Figure, 6.14). Simple main effects were conducted to analyse this interaction. They revealed that for low span subjects, the 15% speech interference in the durable conditions is highly significant (F (1, 7) = 28.974; $p < .01$)[58]. For the transient conditions, the 4% speech interference only approaches significance (F (1, 7) = 3.943; $p < .09$) [59]. The absence of a significant triple interaction between durability, multimodality and span for the complex sentences might be a matter of experimental power, affected by the limited number of span subjects in each complexity condition. However, in its current form, the effect needs to be rejected (see also footnote 59).

---

[58] A One-Sample T-Test revealed that for low span subjects, the comprehension of complex sentences in the dynamic-durable multimodal condition did not depart from a chance level of performance ($t_7 = 1.183$).

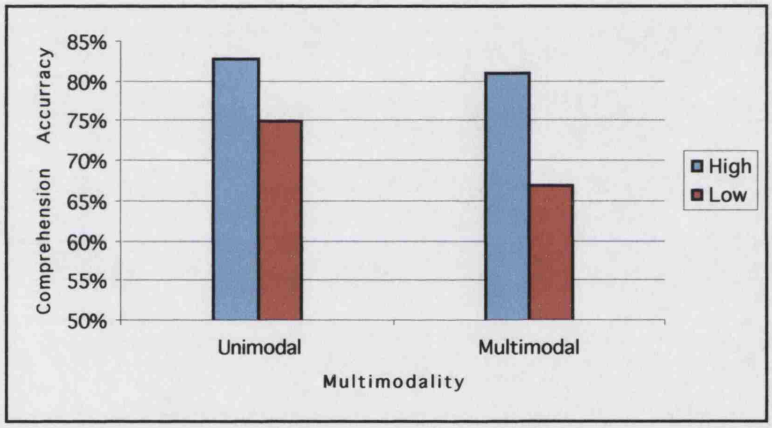
[59] In the version-based analysis, an exploration of the approaching significance interaction between span, durability and multimodality in the complex condition revealed that for high span subjects, the only effect that reached significance was of multimodality (F (1, 4) = 8.115; $p < .05$), (see Figure, 6.13). For low span subjects, both the effect of durability and the effect of multimodality reached significance ($F_{durability}$ (1, 4) = 16.644; $p < .05$; $F_{multimodality}$ (1, 4) = 55.048; $p < .01$). Furthermore, the interaction between these variables approached significance (F (1, 4) = 6.454; $p < .07$), (see Figure, 6.14). Simple main effects revealed that for low span subjects, the 15% speech interference in the durable conditions was highly significant (F (1, 4) = 25.510; $p < .01$). For the transient conditions, the 4% speech interference also reached significance (F (1, 4) = 13.886; $p < .05$). Note that the indication of practice effect for low span subjects that was reported above makes it more difficult to accept the approaching significance interaction. More subjects are needed to validate the assumed relationships between linguistic complexity, durability and verbal WM capacity in the multimodal domain.

Overall, the interaction between span and multimodality reached significance ($F(1, 28) = 4.257$; $p < .05$). The comprehension rates created by these variables are shown in Table 6.12 and Figure 6.15.

Table 6.12
Experiment 2b. Mean Comprehension Rate (CR) for Span Conditions (%): By Multimodality (Standard Errors in Parentheses)

	Multimodality		
Span	Unimodal	Multimodal	Mean difference
High	83%	81%	2%
	(3%)	(2%)	
Low	75%	67%	8%
	(3%)	(2%)	
Mean difference	8%	14%	

Figure 6.15
Experiment 2b. Mean Comprehension Rate (CR) for Span Conditions (%): By Multimodality



Simple main effects demonstrated that the source of the effect is the redundancy cost found for low span subjects ($F(1, 14) = 17.157$; $p < .01$). For high span subjects no effect of multimodality was identified ($F(1, 14) = .961$).

In conclusion, the global pattern of results does not suggest that individual differences in verbal WM capacity are more apparent as processing demands increase (due to the combination of the multimodal presentation technique and the linguistic complexity of the presented materials). For low span subjects, a significant speech interference effect can be observed when processing long-simple sentences involving an immediate assignment of thematic roles, when these are presented in a dynamic-transient form. For complex sentences, the span variable fails to distinguish between varying magnitudes of interference for transient and durable multimodal modes of presentation. More subjects are needed to validate the assumed relationships between linguistic complexity and verbal WM capacity in the multimodal domain.

Response times: The time it took to comprehend each sentence was collected for each subject. A repeated-measure analysis of variance was conducted for the response time data[60]. Complexity and span formed the between-subjects variables while durability and multimodality formed the within-subjects variables. Results suggest that no effect reached significance.

Item Analysis

Comprehension rates were collected for each sentence. Complexity and span formed the within-items variables while durability and multimodality formed the between-items variables. The pattern of the collected data did not warrant the use of parametric tests[61]. However, as suggested earlier, non-parametric tests cannot control for the pragmatic complexity value of an item. Thus, in spite of the fact that the assumption of normality was not met as required, it was decided to perform the item analysis using parametric tests. A repeated-measure analysis of variance was thus conducted, using the pragmatic complexity variable as a covariate. The analysis yielded the following effects: a significant effect of pragmatic complexity ($F(1, 75) = 16.116$; $p < .01$) and a significant interaction between pragmatic complexity and (syntactic) complexity ($F(1, 75) = 8.057$; $p < .01$). The source of this interaction was investigated separately by means of linear regressions. A significant negative relationship was found between pragmatic complexity and comprehension rate for both complex and simple sentences: the higher the pragmatic complexity value, the lower the comprehension. This effect was stronger for complex sentences than for simple sentences.

Predicted CR_{simple} = (.964 - .0013 x Pragmatic complexity + error)
($F(1, 79) = 7.818$; $p < .01$)
Predicted $CR_{complex}$ = (1.034 - .0037 x Pragmatic complexity + error)
($F(1, 79) = 15.796$; $p < .01$)

None of the other effects reached significance.

---

[60] An exploration of the data revealed that for the sub-conditions created by the durability, multimodality and complexity variables and for the sub-conditions created by the durability, multimodality and span variables, the assumption of normality was fully met. Furthermore, the assumption of homogeneity of variance was also met for all conditions.

[61] The exploration of the items' comprehension rate data revealed that the assumption of normality was not met for all sub-conditions created by durability, complexity and span and by multimodality, complexity and span. The assumption of homogeneity of variance was met for all conditions bar the low span complex condition. When squared values were used, they failed again to meet the assumption of normality. On the other hand, the assumption of homogeneity of variance was met for all sub-conditions. It was decided to analyse the squared values of the comprehension rates rather than the original values. Note that the reported mean values were converted back to the original units using a square root transformation.

6.3 General Discussion

6.3.1 Re-examination of the durability account

Adding speech to a dynamic-durable visual text created a multimodal presentation that enabled the user to visually regress to earlier sentential components while simultaneously attending to the spoken continuation of the sentence. This is precisely the pattern of behaviour suggested by the durability account for the multimodal facilitation in comprehending complex sentences that was observed in experiment 1. However, the overall pattern of results obtained in experiment 2 does not show a repeat of the facilitation, nor does it support this account. On the contrary, the interaction found in experiment 2b between complexity, durability and multimodality suggests that speech interference is strongest for complex sentences presented in a dynamic-durable form.

Does this suggest that the multimodal facilitation in comprehending complex sentences, observed in experiment 1, is an artificial finding? Another possible explanation may account for the opposite pattern of results identified in experiment 2: the processing pace of visual information in experiment 2 was machine-driven; visual text was dynamically presented in all presentation conditions. Since dynamic media attracts attention automatically, regressive eye-movements might have been more difficult during the dynamic-durable multimodal presentation used in experiment 2 than during the static-durable multimodal presentation used in experiment 1(that is, in spite of the fact that the slower presentation rate in experiment 2 could have made them theoretically more easy). It is possible that a cross-modal assignment of thematic roles can be conducted only under a static-durable multimodal presentation, in which processing of the visual information is user-paced.

The next experiment aims to clarify the validity of this assertion and to assess the role of *dynamism* in a durable multimodal presentation of sentences that vary in their syntactic complexity. Two modes of durable presentation will be used: a dynamic-durable and a static-durable multimodal presentation. If a cross-modal assignment of thematic roles can be conducted only under user-paced processing of visual information, one would expect to obtain multimodal facilitation in comprehending complex sentences presented in a static-durable form. On the other hand, if increasing sentence complexity impairs the SAS ability to supervise such a sophisticated coordination of processing, one would obtain a strong multimodal interference effect in comprehending complex sentences presented in a static-durable form.

6.3.2 Revised assumptions

The overall pattern of results obtained in experiment 2 suggests that one fundamental assumption made by the MMUM needs to be revised. Specifically, it is proposed that the assumption that the SAS supervises coordination by *synchronisation* between the visual and the auditory channels so as to maximise the multimodal activation of cross-modal sub-systems is over simplified. Using fully coupled multimodal presentations, the experiment identified no evidence of facilitation in

comprehending long-simple sentences. Furthermore, for low span subjects, a significant speech interference effect was found in comprehending long-simple sentences presented in a dynamic-transient form. Although more subjects are needed to validate a similar effect for the dynamic-durable presentation form, this provides some indication that a simultaneous multimodal activation of the cross-modal sub-systems can impair performance even when the assignment of thematic roles is immediate. Moreover, the evidence that multimodality in itself impairs comprehension of complex sentences for all span subjects suggests not only that a simultaneous multimodal activation of the cross-modal sub-systems is negligible under high processing-load conditions, but that the delayed assignment of thematic roles (necessary for the comprehension of complex sentences) is impaired by concurrent speech. Finally, the fact that the strongest interference effect was found in a dynamic-durable multimodal presentation of complex sentences suggests that use of regressive eye-movements exacerbate the magnitude of the impairment in spite of the coupled presentation of the visual and the auditory information.

It seems that no impairment in comprehending sentences presented to both modalities can be attributed solely to the failure of the SAS to supervise coordination by *synchronisation* between modalities. Rather, the coordination between modalities must take a more sophisticated form. Two modes of interference are proposed: one for a dynamic-transient multimodal presentation and the other for a dynamic-durable multimodal presentation. In a dynamic-transient multimodal presentation, multimodal activation of the cross-modal sub-systems *reduces* processing cost at the lexical level. As long as processing load remains low for a particular user, this multimodal activation will optimise information processing in the a-modal sub-systems that are fed by consistent representations of the verified words. When processing load increases, either due to syntactic complexity or (for low span subjects) due to sentence length, processing may breakdown and so users may attempt to retrieve previously processed information. As long as the available resources are sufficient, processing may proceed with no interference. When resources are low, the re-activation of intermediate computational products would be impaired by concurrent speech. The interference might take place at a semantic level of processing, as when background speech impairs performance of a concurrent reading task involving *different* materials (Martin et al., 1988) although an interference on a phonological basis cannot be ruled out.

Similarly, it is suggested that for a dynamic-durable multimodal presentation, multimodal activation of the cross-modal sub-systems reduces user cost as long as sentence load remains low for a particular user. When sentence load increases, users may perform regressive eye-movements to reactivate intermediate computational products while attending to the spoken continuation of the sentence. As long as the available resources are sufficient, processing may proceed with no interference. When resources are low, the recollection of this visual information would be impaired by the concurrent speech, due to the failure of the SAS to supervise coordination of processing between visual and auditory information. (Note that the absence of an interference effect for low span subjects in comprehending long-simple sentences makes it more difficult to qualify when resources are low. At this stage of research, it is clear that resources are low when processing long-complex sentences).

In conclusion, the SAS may supervise coordination by *synchronisation* between the visual and auditory channels so as to maximise the multimodal activation of cross-modal sub-systems as long as processing load remains low for a particular user. When processing load increases, the SAS may use other strategies. In a dynamic-transient multimodal presentation, the SAS may attempt to coordinate between the "real-time" multimodal output and the "forgotten" intermediate computational products. In a dynamic-durable presentation, the SAS may attempt to coordinate between the auditory information and previously processed visual information using regressive eye-movements.

The next experiment should determine whether a dynamic-durable multimodal presentation is indeed superior to a static-durable multimodal presentation in processing simple and complex sentences. Following the production of regressive eye-movements, synchronous processing is easier to restore in a dynamic-durable multimodal presentation through refocusing attention on the "leading edge" of the visual display. If, as suggested by the MMUM, some form of synchronised processing is indeed desirable at high load processing points of the sentence, one would expect a stronger interference in comprehension of complex sentences presented in a non-coupled form relative to those presented in a coupled form to both modalities.

Chapter 7

Experiment 3: the effect of the dynamism of a durable visual text on comprehension of simple and complex sentences presented to both modalities

Experiment 3 investigated the combined effect of the dynamism of durable visual text and the coupling between modalities on user cost, given variations in syntactic complexity. The dynamism of the visual text was systematically varied using two durable presentation techniques: a static-durable format in which the full text appeared simultaneously on the screen and a dynamic-durable format in which words accumulated on the screen to form a sentence. Multimodal presentation consisted of presenting the two visual text conditions with added speech. The dynamic-durable multimodal condition included a coupled presentation of the visual and the spoken words, whereas for the static-durable visual text, the addition of speech is inherently non-coupled (visual text will only be seen as coupled with speech when that text is presented to the reader dynamically).

Results of experiment 3 indicate that multimodality impairs comprehension in all conditions. In addition, in contrast to the predictions made by the MMUM, the results show that regardless of sentence complexity, multimodality and verbal WM capacity, a static-durable presentation is superior to a dynamic-durable presentation. Finally, the results demonstrate a significant interaction between sentence complexity, verbal WM capacity and multimodality. The source of this significant interaction is the stronger resistance of high capacity subjects to speech interference in the simple condition relative to low capacity subjects. In the complex condition, the capacity variable failed to distinguish between comprehension rates in the unimodal and the multimodal conditions or between magnitudes of interference in the dynamic-durable and the static-durable multimodal modes of presentation.

These results necessitate modifying some core assumptions made by the MMUM. Nevertheless, they support the assumption that the SAS relies upon the same limited pool of resources used for sentence processing for its supervision functions. When processing load increases, either due to syntactic complexity or (for low span subjects) due to sentence length, fewer resources are available to the SAS. Consequently, its ability to supervise the coordination of processing between modalities is impaired. The result is a significant multimodal interference effect.

7.1 Introduction

7.1.1 Examination of the extended durability account

The third study aims to examine the effect of coupling spoken and visual words on user cost, with variations in the dynamism of a durable visual text and in syntactic complexity. With the same aim of informing the design of systems wherein users can successfully comprehend various texts with a minimum of processing cost, this study incorporates the previously mentioned principles for the assessment of user cost. The assessment of user cost takes into account the linguistic complexity of the presented materials, the memory demands incurred by their multimodal presentation format and also the verbal WM capacity of the user.

The assessment of user cost in the two previous chapters revealed that multimodality affects sentence comprehension in a differential manner, depending on the syntactic complexity of the sentence and the dynamism of the visual text. In experiment 1, where static-durable visual text was used, it was found that whereas for simple sentences the addition of text improves performance relative to the speech condition, for syntactically complex sentences multimodal presentation improves comprehension relative to both the speech and the visual text conditions. Two possible explanations were raised to account for the multimodal facilitation in comprehending *complex* sentences. A substantive explanation, the durability account, suggests that the availability of the durable visual presentation for a further recollection of information enabled subjects to visually regress to earlier sentential components while simultaneously attending to the spoken continuation of the sentence. This enabled them to advantageously switch attention between modalities so as to perform a delayed assignment of thematic roles across them. This account implies that the resources used by the SAS to supervise the coordination of information between modalities are independent of the resources used for sentence processing. Thus, the SAS is capable of monitoring the performance of the language processing system and of supervising the coordination of information between modalities regardless of sentence complexity. This refutes a central assumption made by the MMUM; the MMUM assumes that the SAS relies upon the same limited pool of resources used for sentence processing for its supervision functions. Thus, increasing sentence complexity increases demand for resources by the language processing system, which therefore reduces the resources available to the SAS. Consequently, the ability of the SAS to supervise the coordination of processing between modalities is impaired.

A second methodological account was raised to explain the multimodal facilitation in comprehending complex sentences observed in experiment 1. This explanation proposes that the observed facilitation was an artifact due to qualitative variations in the monitoring task. It relies on the finding that for both levels of complexity, the word-monitoring task was significantly more difficult in the unimodal conditions relative to the multimodal conditions. If subjects had to allocate more resources to the word-monitoring task in the unimodal conditions, then fewer resources were available for the comprehension task in these conditions. This shortage of resources might have affected the comprehension of simple and complex sentences in a differential manner: whereas it did not affect

the comprehension of the simple sentences, it might have impaired the comprehension of complex sentences in the unimodal conditions relative to the multimodal conditions. Moreover, this methodological explanation suggests that the predictions of the MMUM remain untested. With the removal of the word-monitoring task, the added speech may impair processing of complex sentences, as originally postulated by the MMUM.

Indeed, this central assumption of the MMUM gained support from the results of experiment 2b. Similar to experiment 1, it was found that for both the dynamic-transient and the dynamic-durable visual conditions, the addition of redundant speech did not affect comprehension of simple sentences. However, for both forms of presentation, multimodality severely impaired the comprehension of complex sentences. Specifically, the magnitude of interference was strongest for complex sentences presented in a dynamic-durable form. It was suggested that under a dynamic-durable multimodal presentation of complex sentences, the incompatibility of the spoken information that occurs with visual regressions (following the breakdown of normal processing) produces an interference effect at all levels of processing.

An extension of the durability account could address this apparent inconsistency (i.e., the facilitation in comprehending complex sentences found for the static-durable multimodal format vs. the interference in comprehending complex sentences found for the dynamic-durable multimodal format). Multimodal facilitation in comprehending complex sentences might be limited to the static-durable presentation format in which processing of the visual information is user-paced. In other words, it is possible that a cross-modal assignment of thematic roles can be conducted only under user-paced multimodal processing. Processing visual information in experiment 2b was machine-driven; visual text was dynamically presented in all presentation conditions. Since dynamic media attracts attention automatically, regressive eye-movements might have been difficult to carry out in the dynamic-durable multimodal condition (in spite of the lower presentation rate in that experiment), explaining subjects' failure to assign thematic roles across modalities. In the absence of this constraint, speech would be expected to facilitate the comprehension of complex sentences presented visually.

Note that this extended account challenges the assumption of the MMUM that the SAS relies upon the same limited pool of resources used for sentence processing for its supervision functions. According to this explanation, increasing sentence complexity does not impose demands on the resources used by the SAS, given that processing of the visual information is user-paced. Demands are imposed only when the processing of the visual information is machine-paced. It is the aim of this experiment to examine the validity of this explanation using the static-durable multimodal presentation and the dynamic-durable multimodal presentation. With a static-durable multimodal presentation, visual text has a static form since a whole sentence is presented simultaneously on the screen; processing of the visual text is therefore user-paced. With a dynamic-durable multimodal presentation, visual words appear one at the time (at the same rate of the auditory words) but accumulate on the screen to form a sentence; processing of the visual text is therefore machine-paced. Comparing these two multimodal techniques with their solely visual counterparts will enable the

determination of the role of visual-processing control in processing multimodal information and by so doing, to decide between the extended durability account and the methodological account of the results of experiment 1.

7.1.2 Conflicting assumptions made by the MMUM: memory demands of different visual-presentation techniques vs. the importance of synchronous processing

According to the MMUM, the static-durable and the dynamic-durable visual presentation techniques do not involve the same memory demands. Long sentences presented by static-durable visual text may encourage users to allocate resources to sentence computation rather than to storage of intermediate computational products. This repeated de-allocation of storage resources can induce a temporary forgetting by displacement of intermediate products of comprehension. For complex sentences, where storage and computation demands exceed the available capacity, the unavailability of an earlier piece of information will not allow the parser to compute new elements necessary for the comprehension of the sentence. For simple sentences, the forgotten information can be partially recovered by the recollection of the absent information using regressive eye-movements. In contrast, the dynamic-durable visual presentation forces the pace of reading, making regressive eye-movements more difficult to carry out. As noted in the previous chapters, given a fast rate of presentation, regressive eye-movements must be conducted at the expense of processing a later sentential component. This limitation of the dynamic-durable visual presentation format makes its contrast with the static-durable visual presentation format suitable for the assessment of the role of presentational dynamism and thus the control of visual word-processing in comprehending sentences that vary in their syntactic complexity.

Of greater interest, the addition of speech to these two durable visual-presentation techniques enables an assessment of the effect of visual-processing control on user cost in a multimodal presentation, given variations in syntactic complexity. When the visual text is presented in a static-durable form, the addition of speech results in a non-coupled presentation. The users can determine the synchrony in such a non-coupled presentation through coordinating their reading with their listening. This presentation form has the advantages of allowing the user to scan and to skim the visual text, including making regressive eye-movements to previously processed portions of text. However, despite its high level of visual-processing control, the MMUM specifies a synchronisation problem for this form of presentation: whereas for early sentential components, the lexical sub-system is accessed with redundant information, late sentential components will not necessarily make contact on a redundant multimodal basis due to resource constraints. Although the model assumes that the SAS is capable of supervising the coordination of processing of visual and auditory stimuli by bringing "out of phase" stimuli into phase, it also assumes that increasing processing demands will impair its capability to supervise such synchronisation. Furthermore, the use of regressive eye-movements, associated with increasing processing demands, should impair processing. Specifically, the incompatibility of the spoken information with the recollected visual information will produce an interference effect in the cross-modal sub-systems for all users. As a result, conflicting representations will access the a-modal storage space that serves the syntactic and semantic sub-

systems. Significantly, the MMUM assumes that, for increasing processing demands, the SAS will not have sufficient resources available to supervise the competition for activation between the language sub-systems and to oversee the coordination of information between modalities.

On the other hand, the dynamic-durable multimodal format avoids the synchronised processing problem, as identical visual and auditory words are presented together at the same rate. According to the original version of the MMUM, when visual and auditory stimuli are presented together at the same rate, more resources can be allocated to the language system. The multimodal contact in the lexical sub-system enables processing to feed the a-modal sub-systems with consistent representations of the verified words. Given the assumption that a common pool of resources serves all language processing sub-systems, this multimodal activation implies that more resources are available to the syntactic and the semantic sub-systems. Furthermore, consistent representations maintained by the phonological sub-system enable post-interpretative utilisation of phonological information. The results of experiment 2 demonstrate however that this assumption is over simplified and that speech interference takes place for demanding materials. Using fully coupled multimodal presentations, experiment 2b failed to identify a significant speech interference effect in comprehending long-simple sentences for low span subjects. However, when the data of the dynamic-transient simple conditions from both experiments 2a and 2b was used, a significant speech interference effect was identified for these subjects. This analysis provides some indication that a simultaneous multimodal activation of the cross-modal sub-systems can impair performance even when the assignment of thematic roles is immediate. For complex sentences, the coupled presentation formats severely impaired performance for all span subjects. It seems that under high processing demands, presenting the visual and auditory words together at the same rate does not eliminate the multimodal interference effect.

Yet, the effect of coupled presentation in *repairing asynchronous processing* remains unknown. According to the MMUM, following the production of regressive eye-movements, synchronous processing is easier to restore in the dynamic-durable multimodal format than in the static-durable multimodal format through refocusing attention on the "leading edge" of the visual display. Despite its lower level of visual-processing control, the assumed recoverability property of the dynamic-durable multimodal format might alleviate processing load in comparison to the static-durable multimodal format that lacks this property. It is not suggested that by restoring the synchronous processing, the multimodal interference effect can be eliminated; rather, the affordance to restore synchronous processing may produce a smaller interference effect in the dynamic-durable multimodal conditions than in the static-durable multimodal conditions.

In this, it is also proposed that the property of a visual-processing control is of minor importance under multimodal presentation conditions. The control over visual processing is assumed to assist performance only in a visual presentation of sentences that involve an immediate assignment of thematic roles between sentential components. It is not expected to assist the comprehension of complex sentences. In a multimodal presentation of either simple or complex sentences, the use of regressive eye-movements that is associated with high processing load would interfere with

processing. Under these circumstances, a reduced interference is assumed to take place when synchronous processing is easier to restore than when it is not. Following this assumption, the experimental predictions are provided next.

7.2 Experimental hypotheses

This section tests the theoretical issues discussed above through detailed predictions for the comprehension rate measure in this experiment. Most of the predictions do not refer to the time it takes subjects to respond to each comprehension statement as, again, it was assumed that except the sentence complexity factor, none of the other experimental factors would affect the time it takes to comprehend each statement. Response times were simply collected as a measure of control against unexpected trade-offs between the speed of responses and their accuracy.

7.2.1 Sentence complexity

The MMUM suggests that the excessive storage demands imposed by the need to maintain several unassigned thematic roles in memory will lead to a breakdown in processing the doubly-embedded sentence structure (c.f., Gibson, 1991). The failure to maintain the syntactic structure of the doubly-embedded sentences in WM is expected to yield lower comprehension rates and slower response times than those found for the right-branching sentences in all presentation conditions and for all subjects. For a full rationale, see Chapter 5, section 5.2.1.

7.2.2 Verbal WM capacity and its relationships with visual dynamism, multimodality and sentence complexity

According to the MMUM, individual differences in verbal WM capacity will affect performance when the combination of the multimodal presentation technique and the linguistic complexity of the presented materials imposes a high processing load on the users. The higher the processing load, the more apparent the difference will be. However, this assumption gained only partial support from the results of experiments 2a and 2b. In experiment 2b, the sentence length used was not capable of distinguishing between the comprehension rates of users with low and high verbal WM capacity for simple sentences; neither high nor low span subjects were affected by the added speech. However, the findings for low span subjects seemed to be affected by the low power of that experiment. It is only when the data of the dynamic-transient simple conditions of both experiments 2a and 2b was combined, that results revealed a significant speech interference effect for low span subjects. Yet, a similar effect for the dynamic-durable simple conditions could not be verified based on the results of experiment 2b. On the other hand, for complex sentences, the span variable failed entirely to statistically distinguish between varying magnitudes of interference for transient and durable multimodal modes of presentation. The limited number of span subjects in each complexity condition in experiment 2b might suggest that abandoning the assumed relationships between linguistic

complexity and verbal WM capacity in the multimodal domain is premature. The above assumption will be retained until more data has been gathered.

In this experiment, a significant effect of verbal WM capacity is expected to be found. Overall, high span subjects are expected to show higher comprehension rates than low span subjects for both simple and complex sentences in all presentation conditions. In addition, consistent with the results of experiment 2b, span is expected to interact with multimodality; regardless of sentence complexity and the dynamism of the visual text, low span subjects will experience a greater speech interference effect than high span subjects. Finally, the span variable is expected to significantly interact with the dynamism and multimodality variables only for complex sentences. The expected pattern of results for simple sentences is described next.

Simple sentences

The MMUM assumes that low span subjects rely greatly on the durability of the visual text when processing long-simple sentences. For these subjects, performance is expected to be significantly higher in the static-durable visual condition, which enables recovery of lost information using regressive eye-movements, than in the dynamic-durable visual condition (see Figure 7.1). In contrast, high span subjects are assumed to have a greater capacity to cope with the storage and computational demands of long simple sentences. For these subjects, a smaller effect of visual dynamism is expected to be found (see Figure 7.2).

Figure 7.1

Low Span Subjects: Predicted User Cost as a function of Presentation Type for Long-Simple Sentences (Revised following the combined analysis of experiments 2a and 2b)

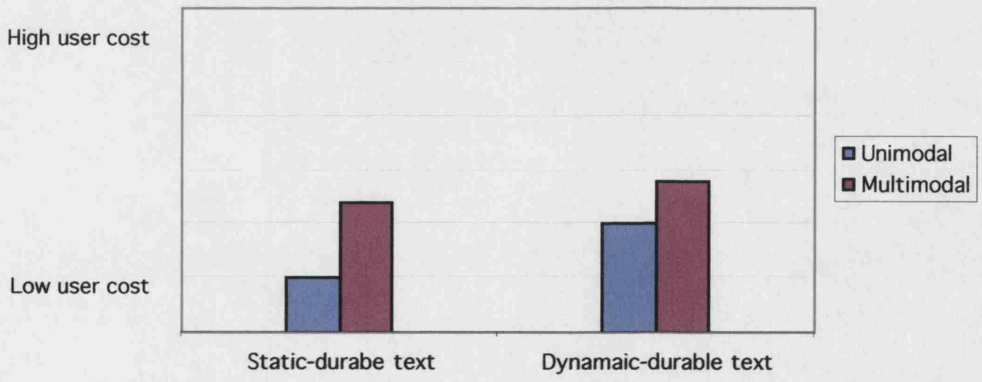
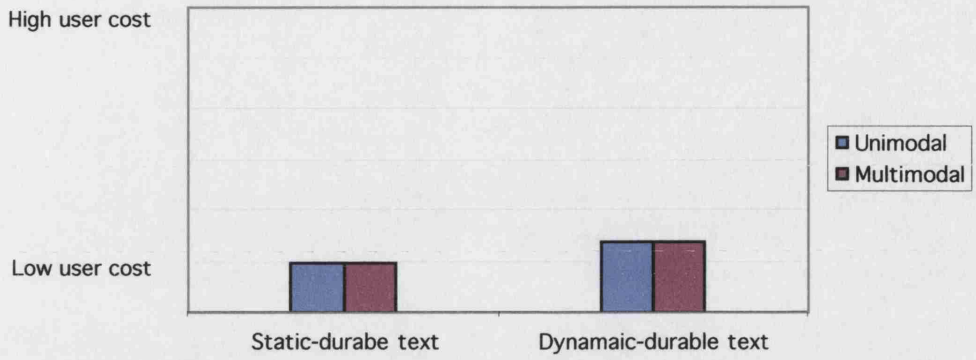


Figure 7.2

High Span Subjects: Predicted User Cost as a function of Presentation Type for Long-Simple Sentences (Revised following the combined analysis of experiments 2a and 2b)



In addition, following the combined analysis of experiments 2a and 2b, it is suggested that multimodality may affect the comprehension rates of low and high span subjects differentially also in durable conditions and that the lack of the effect in the dynamic-durable conditions in experiment 2b was due to the limited number of span subjects in each complexity condition. The length of the right-branching sentences is assumed to impose a high processing load on low span subjects in both static and dynamic durable conditions of this study. Low span users may attempt to make regressive eye-movements to reactivate intermediate computational products while attending to the spoken continuation of the sentence in both multimodal conditions. However, because resources are low, the recollection of the visual information will be impaired by the concurrent speech, due to the failure of the SAS to supervise coordination of processing of visual and auditory information. On the other hand, high span subjects are expected to experience lower processing load when processing these sentences. Their reduced reliance on the durability of the visual text suggests a reduced dependence on regressive eye movements. Finally, their SAS is assumed to have sufficient resources to coordinate both storage and computational demands of long-simple sentences and to maintain a successful divided attention between modalities.

Determining the relationship between span, dynamism and multimodality for simple sentences is more problematic. If the dynamism of the visual text were the only factor to determine processing load, then consistent with the assumed relationship between verbal WM capacity and processing load, one would expect that low span subjects would experience a stronger speech interference effect in the dynamic-durable conditions than in the static-durable conditions, as the dynamic-durable visual condition imposes a greater processing load on these users. Furthermore, this difference would be predicted to be only slightly larger than the expected difference for high span subjects. The experimental predictions rely however on an additional assumption, namely the *recoverability* of synchronous processing in various coupled presentations. As noted earlier, the MMUM assumes that following a regressive eye-movement, synchronous processing is easier to restore in the dynamic-durable multimodal format than in the static-durable multimodal format through refocusing attention on the "leading edge" of the visual display. Thus, the model implicitly prioritises the importance of

the recoverability property of the dynamic-durable multimodal format over the property of visual-processing control of the static-durable multimodal format in alleviating processing cost.

With this prioritisation in mind, it is predicted that low span subjects will experience a greater speech interference effect in the static-durable conditions than in the dynamic-durable conditions. This difference will be only slightly larger than that predicted for high span subjects. Consistent with the results of experiment 2b, it is assumed that high span users will not experience a significant speech interference effect in the dynamic-durable condition. Similarly, for the static-durable condition, these subjects are not expected to experience a significant speech interference effect. For both forms of presentation, their SAS is expected to have sufficient resources available to supervise the competition for activation between the language sub-systems and to oversee the coordination of information between modalities.

In summary, for long-simple sentences, predictions include a main effect of span, a main effect of dynamism (higher performance in the static than in the dynamic conditions), no effect of multimodality and an interaction between span and multimodality (an interference effect of multimodality for low span subjects and no effect of multimodality for high span subjects). No triple interaction between these variables is predicted for long-simple sentences.

Complex sentences

For long-complex sentences, a triple interaction is expected between span, dynamism and multimodality.

The MMUM assumes that for both unimodal conditions, regressive eye-movements will not assist the processing of complex sentences. Consistent with Gibson's (1991) complexity metric and the results of the unimodal conditions in experiment 1, the syntactic parser will not have sufficient resources available for the delayed assignment of thematic roles required for the comprehension of complex sentences for either dynamic or static forms of presentation. The model therefore predicts a main effect of span, no effect of dynamism and no interaction between span and dynamism for complex sentences presented visually.

A different pattern of performance is predicted for multimodal presentation of complex sentences. The static-durable multimodal condition enables the user to flexibly use eye regressions while attending to the spoken continuation of the sentence. Again, this is precisely the pattern of behaviour suggested by the extended durability account for the multimodal facilitation in comprehending complex sentences that was observed in experiment 1. If this account is valid and a static visual presentation does enable the SAS to supervise the coordination of information between modalities, speech facilitation should occur in comprehension of the doubly-embedded sentences in the static-durable conditions. Moreover, this facilitation will be larger for high than for low span subjects. Since regressive eye-movements are difficult to carry out in the dynamic-durable conditions, all subjects will fail to assign thematic roles across modalities and will demonstrate a strong speech

interference effect in these conditions. Finally, this interference will be greater for low than for high span subjects.

However, the MMUM provides a different prediction. According to the model, the use of regressive eye-movements while attending to the spoken continuation of an excessively complex sentence will be an unsuccessful strategy. Specifically, the incompatibility of the spoken information with the recollected visual information will produce an interference effect in the cross-modal sub-systems for all users. As a result, conflicting representations will access the a-modal storage space that serves the syntactic and semantic sub-systems. The MMUM assumes that all users' SAS will not have sufficient resources available to supervise the competition for activation between the language sub-systems and to oversee the coordination of information between modalities. A strong speech interference effect is therefore predicted for the complex sentences. Moreover, assuming that synchronous processing is easier to restore in the dynamic-durable coupled format than in the static-durable non-coupled format through refocusing attention on the "leading edge" of the visual display, subjects will experience a greater speech interference effect in the static-durable than in the dynamic-durable conditions. Finally, since low span subjects are assumed to have a smaller verbal WM capacity than high span subjects, a significantly greater interference effect is expected to be found for these subjects in comprehending long-complex sentences in the static-durable multimodal condition. This difference will be larger than that predicted for the dynamic-durable multimodal presentation (see Chapter 4, Figures 4.6 and 4.7).

In summary, for long-complex sentences, predictions include: no effect of dynamism, a main effect of span, a main effect of multimodality, an interaction between span and multimodality (a greater speech interference effect for low span subjects than for high span subjects), an interaction between dynamism and multimodality (a greater interference in the static-durable conditions than in the dynamic-durable conditions) and a triple interaction between dynamism, multimodality and span.

On average, and regardless of users' verbal WM capacity, the MMUM predicts a significant triple interaction between complexity, dynamism and multimodality. The source of this triple interaction is the expected interaction between dynamism and multimodality for the complex condition. As noted earlier, for simple sentences these variables are not expected to interact. For complex sentences, speech interference will be larger in the static-durable multimodal condition than in the dynamic-durable multimodal condition, assuming that the latter enables synchronous processing to be restored through refocusing attention on the "leading edge" of the visual display.

Overall, this pattern of behaviour should lead to a significant interaction between dynamism and multimodality. Synchronous processing is assumed to be easier to restore in the dynamic-durable multimodal condition regardless of sentence complexity; a reduced interference effect is therefore expected to be found in this condition relative to the static-durable multimodal condition. Results are also expected to yield a significant interaction between complexity and dynamism: whereas for simple sentences, any absent information can be recovered using regressive eye-movements, for complex sentences this will not prove useful. Consistent with the results of experiment 2b, a

significant interaction is expected to be found between complexity and multimodality: a weak interference of multimodality in the comprehension of long-simple sentences and a strong interference of multimodality in the comprehension of long-complex sentences. Finally, this pattern of results is expected to produce a significant effect of multimodality; speech is expected to impair comprehension in this experiment.

7.3 Method

7.3.1 Materials and design

The primary stimulus set consisted of the same 80 pairs of sentences used in experiment 2, each containing one right-branching sentence and one doubly-embedded sentence.

Subjects were assigned to one of the two experimental groups defined by the complexity factor[62]. Each complexity group was presented with the sentences in four presentation conditions created by the dynamism and the multimodality factors[63]:
1. Dynamic-durable visual presentation
2. Static-durable visual presentation
3. Dynamic-durable multimodal presentation
4. Static-durable multimodal presentation

The same sound files, created for experiment 2, were used again in this experiment[64]. Also, the visual files, created for the dynamic-durable conditions of experiment 2b, were used in the dynamic-durable conditions of this experiment (all sentences were presented in one line). New visual files were created for the static-durable conditions; each file containing a full sentence presented in one line. For a full description of the materials preparation, see Chapter 6, section 6.1.3.

7.3.2 Comprehension statements

The same comprehension statements, created for experiment 2, were used again in this experiment.

7.3.3 Implementation

General implementation principles of the sentence materials and the comprehension sentences were identical to those used in experiment 2. Four experimental versions were created for each complexity group; each having a different order of the presentation-conditions blocks to minimise order effects[65]:

[62] See Chapter 5, footnote 23.
[63] See Chapter 6, footnote 33.
[64] See Chapter 6, footnote 34.
[65] See Chapter 5, footnote 27.

Version 1: dynamic-durable visual presentation, dynamic-durable multimodal presentation, static-durable visual presentation and static-durable multimodal presentation
Version 2: static-durable visual presentation, static-durable multimodal presentation, dynamic-durable visual presentation and dynamic-durable multimodal presentation
Version 3: dynamic-durable multimodal presentation, dynamic-durable visual presentation, static-durable multimodal presentation and static-durable visual presentation
Version 4: static-durable multimodal presentation, static-durable visual presentation, dynamic-durable multimodal presentation and dynamic-durable visual presentation

7.3.4 Apparatus

The experiment was run on a PowerPC G3/300. Spoken sentences were presented at a comfortable volume through headphones (Vivanco SR 250) and the visual material was presented on a Compaq display, size 17 inches. Subjects responses to the comprehension statements and the times of these responses were collected via Apple Desktop Bus™ (ADB) keyboard (see experiment 1).

7.3.5 Procedure

Identical to experiment 2.

7.3.6 Span task

Subjects' verbal WM capacity was measured in a separate session. The same testing procedure was followed as in the previous experiments. 16 subjects whose reading spans were 3.5 or higher were classified as high span subjects and 16 subjects whose reading spans were 3.0 or lower were classified as low span subjects.

7.3.7 Subjects

32 volunteers and students at University College London were tested and were paid £6 for their participation. English was the first language of all subjects.

7.4 Results & Discussion

7.4.1 Comprehension rates

Subjects' analysis of overall rates

Comprehension rates were collected for each subject and were analysed in a repeated-measure analysis of variance. Complexity and span formed the between-subjects variables while dynamism and multimodality formed the within-subjects variables[66],[67]. The comprehension rates of simple sentences (89%) were 22% higher than those found for complex sentences (67%), yielding the expected main effect of complexity ($F(1, 28) = 66.526$; $p < .01$). In addition, comprehension rates in the multimodal conditions (74%) were 8% lower than those in the unimodal conditions (82%), yielding the expected main effect of multimodality ($F(1, 28) = 53.354$; $p < .01$). However, the overall pattern of results is not compatible with the predictions made for this experiment. Contrary to the experimental predictions and the results of experiment 2b, the analysis failed to produce a significant interaction between complexity and multimodality ($F(1, 28) = .243$). Table 7.1 presents the mean comprehension rates created by these variables. The relationship between these means can be seen more readily in Figure 7.3.

---

[66] An exploration of the comprehension rate data revealed that the sub-conditions created by complexity, dynamism and multimodality did not exhibit perfectly normal distributions. The assumption of normality was not met for simple sentences in the dynamic-durable visual condition (Shapiro-Wilk (16) = .871; $p < .04$), in the static-durable visual condition (Shapiro-Wilk (16) = .848; $p < .02$) and in the static-durable multimodal condition (Shapiro-Wilk (16) = .867; $p < .04$). Bar one condition, all distributions were positively skewed. In addition, the assumption of homogeneity of variance was not met for both the dynamic-durable visual condition and the static-durable visual condition. Moreover, the sub-conditions created by span, dynamism and multimodality did not exhibit perfectly normal distributions. Specifically, the assumption of normality was not met for low span subjects in the static-durable multimodal condition (Shapiro-Wilk (14) = .873; $p < .04$), for high span subjects in the dynamic-durable visual condition (Shapiro-Wilk (16) = .841; $p < .01$), and in the static-durable visual condition (Shapiro-Wilk (16) = .875; $p < .05$). Finally, all distributions were negatively skewed. An additional exploration used the squared values of the comprehension rate data. Whereas this transformation did not affect the normality values, the assumption of homogeneity of variance was now met for three presentation conditions created by the complexity, dynamism and multimodality variables. Furthermore, the probability of meeting the assumption of homogeneity of variance for the dynamic-durable visual condition improved. It was decided to conduct the analyses of variance over the transformed values of the comprehension rate measure rather than the original values. Note that the reported mean values were converted back to the original units using a square root transformation.

[67] An additional analysis was conducted at the request of the examiners of this dissertation. This analysis included the order in which subjects performed the four presentation conditions as an additional between-subjects variable (called version number) to make sure that the experimental results were unaffected by order effects (see Method, section 7.3.3). None of the results reported in this section was affected by the order in which subjects performed the four presentation conditions. Note that the number of subjects in each version is too small to make this a reliable conclusion.

Table 7.1

Experiment 3. Mean Comprehension Rate (CR) for Complexity Conditions (%): By Multimodality (Standard Errors in Parentheses)

	Multimodality		
Complexity	Unimodal	Multimodal	Mean difference
Simple	93%	86%	7%
	(2%)	(2%)	
Complex	72%	62%	10%
	(2%)	(2%)	
Mean difference	21%	24%	

Figure 7.3

Experiment 3. Mean Comprehension Rate (CR) for Complexity Conditions (%): By Multimodality

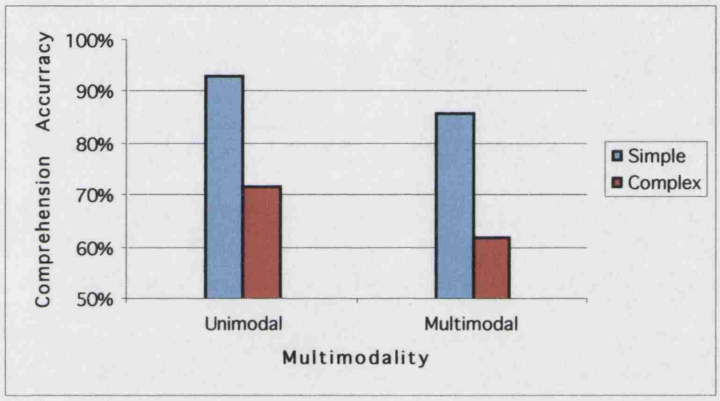


Figure 7.3 shows that for both complexity conditions, multimodality impairs performance. Whereas for complex sentences this result was anticipated, multimodality was not expected to significantly affect comprehension of simple sentences. An inspection of the comprehension rates of low span subjects suggests that their poor performance in the multimodal conditions accounts for the overall speech interference effect in comprehending simple sentences (see Table 7.6). The varying comprehension rates of low and high span subjects and their effect on the overall pattern of results will be explained in the next section, concerning individual differences in verbal WM capacity.

Results also failed to validate the assumed relationship between complexity and the dynamism of the visual text. Table 7.2 and Figure 7.4 present the mean comprehension rates created by these variables.

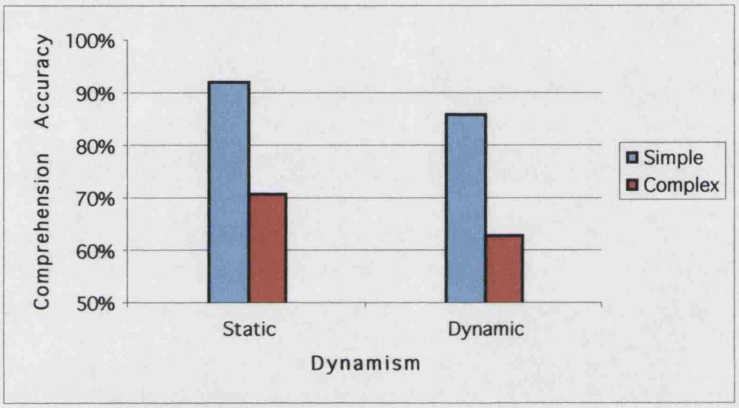
Table 7.2

Experiment 3. Mean Comprehension Rate (CR) for Complexity Conditions (%): By Dynamism (Standard Errors in Parentheses)

	Dynamism		
Complexity	Static	Dynamic	Mean difference
Simple	92%	86%	6%
	(2%)	(3%)	
Complex	71%	63%	
	(2%)	(3%)	8%
Mean difference	21%	23%	

Figure 7.4

Experiment 3. Mean Comprehension Rate (CR) for Complexity Conditions (%): By Dynamism



Contrary to expectation, Figure 7.4 shows that durability and complexity did not interact. ($F(1, 28) = .922$). Rather, performance with static text is improved in both complexity conditions (82% in the static conditions Vs. 75% in the dynamic conditions), ($F(1, 28) = 36.897$; $p < .01$). For simple sentences this result was predicted; performance was expected to be significantly higher in the static-durable conditions that enable the user to recover lost information using regressive eye-movements than in the dynamic-durable conditions in which regressive eye-movements are more difficult to carry out. Furthermore, this result is consistent with the finding of experiment 1 that for simple sentences, performance in the static-durable visual text condition was higher than in the dynamic-transient speech condition. On the other hand, for complex sentences, the dynamism of the visual text was not expected to affect comprehension. Consistent with Gibson's (1991) complexity metric, it was assumed that for excessively complex sentences, the syntactic parser would not have sufficient resources available for the delayed assignment of thematic roles regardless of the dynamism of the visual text[68]. The effect of dynamism found for complex sentences contradicts this assumption.

[68] Specifically, the dynamism of the visual text was not expected to affect comprehension rates of highly complex sentences in the unimodal conditions. Furthermore, the advantage predicted for the dynamic-durable

Moreover, it is also incompatible with the finding of experiment 1 that for complex sentences, comprehension rates in the static-durable visual text condition were equal to those found for the dynamic-transient speech condition.

It is suggested that differences in sentence presentation rate in the two experiments can address this inconsistency. In experiment 1, presentation duration of each sentence was on average 4074 ms (204 words per minute). Although relatively high, it was assumed that a slower presentation rate would allow subjects to visually scan for late target words in the simple visual conditions (see Chapter 5, section 5.3.1). Presentation duration in experiment 3 was on average 4990 ms (159 words per minute). The additional 916 ms per sentence in this experiment did not prevent the parse from failing, yet the extra-time might have facilitated the repair process of the syntactic breakdown in the static-durable conditions. This implies that under slow presentation rate conditions, a higher visual-processing control facilitates the assignment of thematic roles between sentential constituents. Under dynamic-durable presentation conditions, the poor visual-processing control does not enable the repair of syntactic breakdown to the same extent.

The results also failed to validate the predicted relationship between complexity, dynamism and multimodality. Whereas for simple sentences dynamism and multimodality were not expected to interact, it was assumed that for complex sentences, speech interference would be stronger in the static than in the dynamic conditions. However, sentence complexity failed to distinguish between varying magnitudes of multimodal interference for these modes of presentation ($F(1, 28) = .824$). Table 7.3 and Figure 7.5 demonstrate the relationships between dynamism and multimodality for simple sentences. Table 7.4 and Figure 7.6 demonstrate these relationships for complex sentences.

Table 7.3

Experiment 3. Mean Comprehension Rate (CR) for the Simple Conditions (%): By Dynamism and Multimodality (Standard Errors in Parentheses)

	Multimodality		
Dynamism	Unimodal	Multimodal	Mean difference
Static	95%	90%	5%
	(2%)	(2%)	
Dynamic	91%	82%	9%
	(3%)	(3%)	
Mean difference	4%	8%	

multimodal condition over the static-durable multimodal condition was not expected to be large enough to establish an effect of dynamism for complex sentences when calculated across both multimodality conditions.

Figure 7.5

Experiment 3. Mean Comprehension Rate (CR) for the Simple Conditions (%): By Dynamism and Multimodality

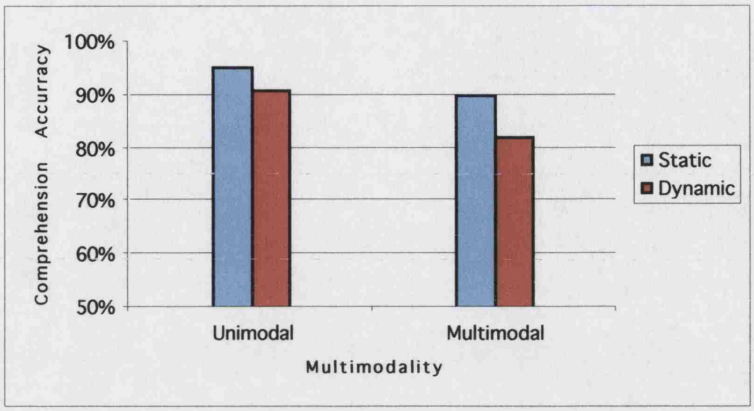


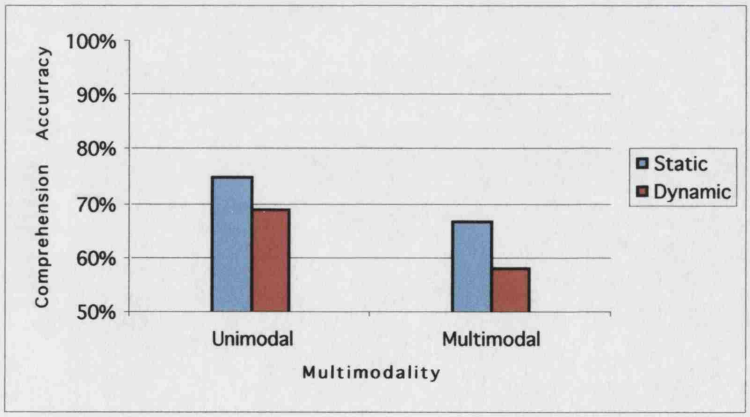
Table 7.4

Experiment 3. Mean Comprehension Rate (CR) for the Complex Conditions (%): By Dynamism and Multimodality (Standard Errors in Parentheses)

	Multimodality		
Dynamism	Unimodal	Multimodal	Mean difference
Static	75%	67%	8%
	(2%)	(2%)	
Dynamic	69%	58%	9%
	(3%)	(3%)	
Mean difference	6%	9%	

Figure 7.6

Experiment 3. Mean Comprehension Rate (CR) for the Complex Conditions (%): By Dynamism and Multimodality



Consistent with the experimental predictions, dynamism and multimodality did not interact for the simple sentences ($F(1, 14) = 1.11$). However, contrary to the predictions made by the MMUM, the effects of dynamism and multimodality, reported earlier for the complex sentences, were also additive and failed to interact ($F(1, 14) = .302$). Figure 7.6 demonstrates that speech interferes with the comprehension of complex sentences regardless of the dynamism of the multimodal format. In addition, the figure demonstrates that the facilitation of visual-processing control in comprehending complex sentences is manifested in both the unimodal and multimodal conditions. Finally, a One-Sample T-Test revealed that comprehension rates of complex sentences in the dynamic-durable multimodal condition did not depart from a chance level of performance ($t_{15} = 2.139$). The superior performance in the static-durable multimodal condition over the dynamic-durable multimodal condition suggests that the importance of a coupled presentation in helping the user to restore synchronous processing was overestimated by the MMUM. The coupling of the "leading edge" of the visual display with the spoken words does not improve comprehension of complex sentences relative to a non-coupled presentation of such sentences.

On the other hand, the multimodal interference in the static conditions clearly refutes the extended durability account suggested to explain the multimodal facilitation in comprehending complex sentences in experiment 1. Nevertheless, this finding provides support to the central claim of the MMUM that the SAS relies upon the same limited pool of resources used for sentence processing for its supervision functions. Increased sentence complexity imposes demands not only on the resources that are used by the language processing system, but also on the resources used by the SAS. Consequently, the ability of the SAS to supervise the coordination of processing between modalities deteriorates.

Finally, at this preliminary state of research, the absence of a significant interaction between dynamism and multimodality in comprehending complex sentences seems to complement the finding of a significant interaction between durability and multimodality for complex sentences in experiment 2b (a greater speech interference in the dynamic-durable conditions than in the dynamic-transient conditions). It provides an important qualification regarding the varying magnitudes of multimodal interference in comprehending complex sentences; it suggests that regardless of the dynamism of the visual text, speech interference is stronger following regressive eye-movements than following the recollection of intermediate computational products in a transient multimodal presentation.

Overall, the pattern of comprehension rates in the complex conditions requires a significant modification of the MMUM: it is suggested that under a durable multimodal presentation of complex sentences, users make regressive eye-movements to reactivate intermediate computational products while attending to the spoken continuation of the sentence. Because resources are low, the recollection of this visual information is impaired by the concurrent speech, due to the failure of the SAS to supervise coordination of processing of visual and auditory information. Moreover, this speech interference is not affected by the dynamism of the durable visual text. In other words, following regressive eye-movements, the recovery from the interfering asynchronous processing is

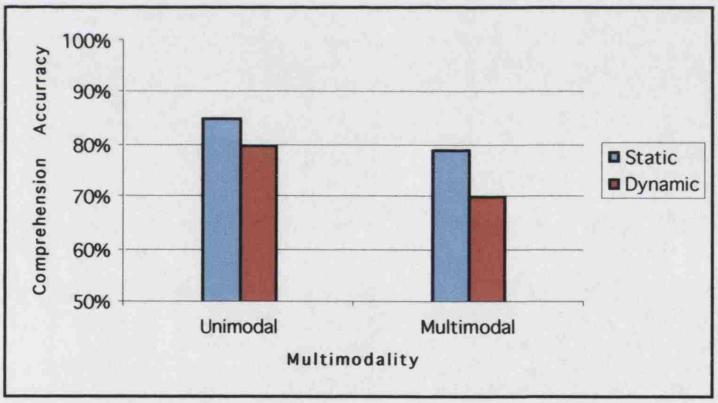
not facilitated by the presentational coupling. Most importantly, the higher visual-processing control in the static-durable multimodal condition facilitates the repair process of the syntactic breakdown, despite the speech interference effect.

Calculated across complexity conditions, the analysis failed to yield a significant interaction between dynamism and multimodality ($F(1, 28) = 1.373$). Table 7.5 and Figure 7.7 demonstrate the overall relationships between dynamism and multimodality in this experiment.

Table 7.5
Experiment 3. Mean Comprehension Rate (CR) (%): By Dynamism and Multimodality
(Standard Errors in Parentheses)

	Multimodality		
Dynamism	Unimodal	Multimodal	Mean difference
Static	85%	79%	6%
	(2%)	(2%)	
Dynamic	80%	70%	10%
	(2%)	(2%)	
Mean difference	5%	9%	

Figure 7.7
Experiment 3. Mean Comprehension Rate (CR) (%): By Dynamism and Multimodality



It appears that a high visual-processing control is a stronger factor than a coupled multimodal presentation; the coupling of the "leading edge" of the visual display with the spoken words does not improve comprehension relative to a non-coupled presentation.

Consistent with the experimental predictions, comprehension rates of high span subjects (81%) were 5% higher than those found for low span subjects (76%), yielding a main effect of span ($F(1, 28) = 5.120$; $p < .04$). The expected relationships between the span variable and the other variables were only partially confirmed by the experimental results. Specifically, for simple sentences presented in a durable form, predictions included a main effect of span and an interaction between span and multimodality: an interference effect of multimodality for low span subjects and no effect of multimodality for high span subjects. Consistent with the results of experiment 2b, high span subjects were assumed to have sufficient resources to coordinate both storage and computational demands of simple sentences and a successful divided attention between modalities. For complex sentences, predictions included a main effect of span, an interaction between span and multimodality (a greater speech interference effect for low span subjects than for high span subjects) and a triple interaction between span, dynamism and multimodality. Specifically, a significantly larger speech interference effect was predicted for low span subjects than for high span subjects in comprehending complex sentences delivered by the static-durable multimodal presentation. This difference was expected to be larger than that predicted for the dynamic-durable multimodal condition.

The analysis failed to confirm this complex pattern of results. Contrary to the experimental predictions, a significant interaction was found between complexity, span and multimodality ($F(1, 28) = 5.830$; $p < .03$). The investigation of this interaction involved a reduced analysis of variance at each level of the complexity variable. Table 7.6 and Figure 7.8 demonstrate the relationships between span and multimodality for the simple sentences.

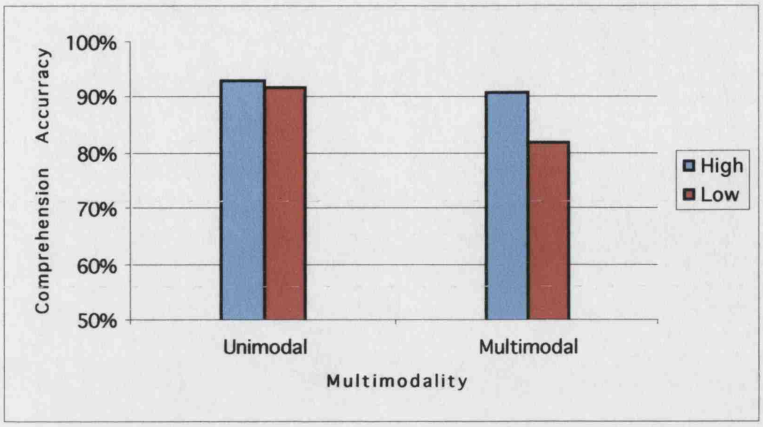
Table 7.6

Experiment 3. Mean Comprehension Rate (CR) for Simple Sentences (%): By Multimodality and Span (Standard Errors in Parentheses)

	Multimodality		
Span	Unimodal	Multimodal	Mean difference
High	93%	91%	2%
	(3%)	(3%)	
Low	92%	82%	10%
	(3%)	(3%)	
Mean difference	1%	9%	

Figure 7.8

Experiment 3. Mean Comprehension Rate (CR) for Simple Sentences (%): By Multimodality and Span



The predictions for the simple sentences were fully validated by the observed comprehension rates. The analysis yielded a significant interaction between span and multimodality ($F(1, 14) = 6.936$; $p < .03$). Simple main effects showed that for low span subjects, the 10% speech interference effect is highly significant ($F(1, 7) = 19.740$; $p < .01$). For high span subjects, concurrent speech did not affect comprehension of simple sentences[69] ($F(1, 7) = 1.430$).

Overall, in the simple condition high span subjects performed significantly better (92%) than low span subjects (87%) ($F(1, 14) = 7.989$; $p < .02$). These results complement the findings obtained in the combined analysis of experiments 2a and 2b for long-simple sentences presented in a dynamic-transient form. They indicate that regardless of the durability or the dynamism of the visual text, long-simple sentences impose varying processing demands on low and high span users. For low span users, long-simple sentences impose a high processing load. It is suggested that although thematic roles can be immediately assigned, low span users fail to retain a full representation of intermediate products of computation in WM (c.f., Miyake et al., 1994). Their attempt to reactivate these computational products (either physically or mentally) is interfered by the concurrent speech; their SAS cannot accommodate both the storage and computation demands of long-simple sentences and the requirement of supervising coordination of processing between modalities. On the other hand, high span subjects experience lower processing load when processing these sentences. As indicated by the absence of a significant interaction between span and dynamism in the simple condition ($F(1, 14) = .383$), high span users also performed better under static-durable than under dynamic-durable

---

[69] Note that although the interaction between dynamism, multimodality and span did not reach significance for simple sentences, the comprehension rates of high span subjects in the dynamic-durable multimodal condition (86%) were lower than their comprehension-rates in the dynamic-durable visual condition (91%). In experiment 2b, speech slightly and insignificantly improved their performance in the dynamic-durable simple conditions. Moreover, the interference found in this experiment contributes to the overall speech interference in the simple conditions. This difference can be attributed to i) the limited potential to screen span subjects in both experiments, and ii) the small number of span subjects in both experiments. These factors are discussed later in this section.

conditions. However, even if these users also used eye regressions to reactivate intermediate products of computation, their SAS had sufficient resources to coordinate both storage and computational demands of long-simple sentences and to maintain a successful divided attention between modalities.

In contrast to the successful predictions for the simple conditions, the comprehension rates of complex sentences failed to validate any of the assumed predictions for the span variable. First, the span variable failed to distinguish between magnitudes of speech interference in the static and the dynamic presentation formats of complex sentences (F (1, 14) = .824). Table 7.7 and Figure 7.9 demonstrate the relationships between dynamism and multimodality for high span subjects. Table 7.8 and Figure 7.10 show the relationships between these variables for low span subjects.

Table 7.7

Experiment 3. Mean Comprehension Rate (CR) of High Span Subjects for Complex Sentences (%): By Dynamism and Multimodality (Standard Errors in Parentheses)

	Multimodality		
Dynamism	Unimodal	Multimodal	Mean difference
Static	80%	70%	10%
	(3%)	(3%)	
Dynamic	71%	61%	10%
	(4%)	(4%)	
Mean difference	9%	9%	

Figure 7.9

Experiment 3. Mean Comprehension Rate (CR) of High Span Subjects for Complex Sentences (%): By Dynamism and Multimodality

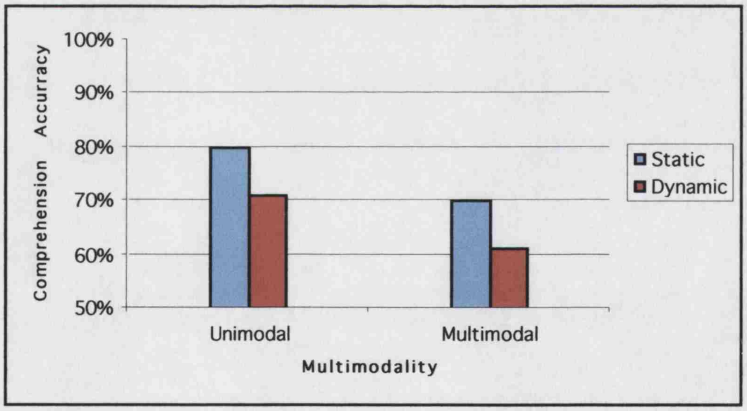


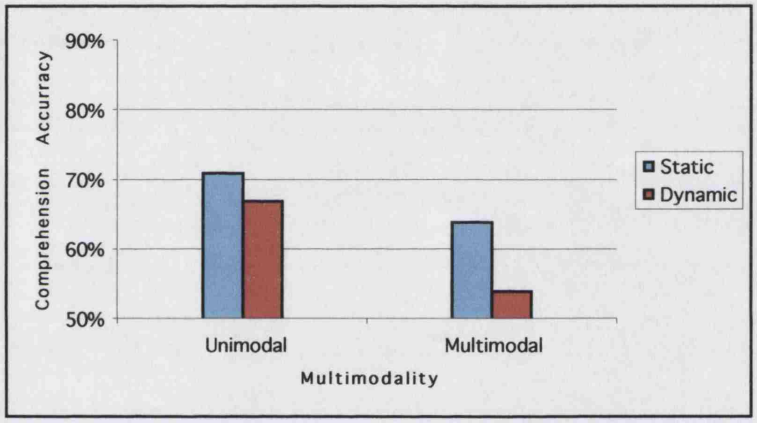
Table 7.8

Experiment 3. Mean Comprehension Rate (CR) of Low Span Subjects for Complex Sentences (%): By Dynamism and Multimodality (Standard Errors in Parentheses)

	Multimodality		
Dynamism	Unimodal	Multimodal	Mean difference
Static	71%	64%	7%
	(3%)	(3%)	
Dynamic	67%	54%	13%
	(4%)	(4%)	
Mean difference	4%	10%	

Figure 7.10

Experiment 3. Mean Comprehension Rate (CR) of Low Span Subjects for Complex Sentences (%): By Dynamism and Multimodality



These figures demonstrate that the absence of a significant interaction between dynamism and multimodality in the complex conditions described earlier, is not mediated by the span variable (see also Figure 7.6). Results clearly show that low span subjects experienced a similar speech interference effect to that of high span subjects in the dynamic-durable conditions (13% vs. 10%). Another peculiar result is that *stronger* speech interference was found for high span subjects than for low span subjects in the static-durable conditions (10% vs. 7%). Although statistically insignificant, this result raises doubts regarding the validity of the screening method of low and high span users. As described earlier, Daneman & Carpenter's (1980) method suggests that subjects whose reading spans are 2.5 or lower should be classified as low span subjects, those whose reading spans are 3 or 3.5 should be classified as medium-span subjects and those whose reading spans are 4 or higher should be classified as high span subjects. Medium-span subjects were to be eliminated from any span-based analysis. Unfortunately, the available resources required using their data as not enough subjects met the definition of low span. Subjects whose reading spans were 3 were classified as low span subjects, whereas subjects whose reading spans were 3.5 were classified as high span subjects. It is possible that this screening method blurred real differences in the performance of the two span groups.

Furthermore, it seems that a floor effect operated in the complex conditions, failing to identify a significant interaction between dynamism, multimodality and span for complex sentences. This suggestion relates to a different pattern of interaction than the one proposed earlier. In the analysis conducted across complexity conditions, a high visual-processing control was found to be more important in reducing processing load than the coupling between modalities, making the dynamic-durable multimodal condition more demanding than the static-durable multimodal condition. Following with the assumed relationship between verbal WM capacity and processing load, a significantly larger speech interference effect should have been found for low span subjects than for high span subjects in comprehending complex sentences delivered by the dynamic-durable multimodal presentation. Moreover, this difference should have been larger than the difference found for the static-durable multimodal condition. In contrast, a One-Sample T-Test revealed that comprehension rates of complex sentences in the dynamic-durable multimodal condition did not depart from chance level of performance for both groups of span (High Span: $t_7 = 1.668$; Low Span: $t_7 = 1.433$). If high span subjects hit the floor in their comprehension of complex sentences in the dynamic-durable multimodal condition, this might have defeated the attempt to reveal a greater speech interference effect for low span subjects in this condition. Consequently, a three-ways interaction could not have been identified for complex sentences.

Whilst accepting these interpretation constraints, the data supports the claim that under a durable multimodal presentation of excessively complex sentences, users use regressive eye-movements to reactivate intermediate computational products while attending to the spoken continuation of the sentence. Because resources are low for both low and high span subjects, the recollection of this visual information is impaired by the concurrent speech, due to the failure of the SAS to supervise coordination of processing of visual and auditory information.

Table 7.9 and Figure 7.11 demonstrate that the span variable also failed to distinguish between magnitudes of speech interference for complex sentences when calculated across the dynamism variable ($F(1, 14) = .037$).

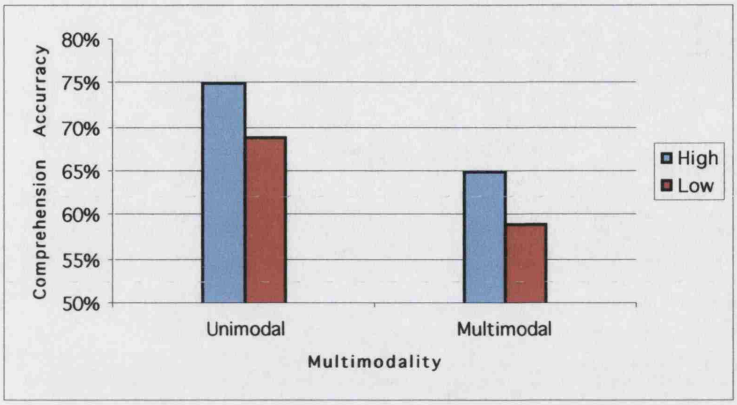
Table 7.9

Experiment 3. Mean Comprehension Rate (CR) for Complex Sentences (%): By Multimodality and Span (Standard Errors in Parentheses)

	Multimodality		
Span	Unimodal	Multimodal	Mean difference
High	75%	65%	10%
	(3%)	(3%)	
Low	69%	59%	10%
	(3%)	(3%)	
Mean difference	6%	6%	

Figure 7.11

Experiment 3. Mean Comprehension Rate (CR) for Complex Sentences (%): By Multimodality and Span



Finally, although high span subjects achieved higher comprehension rates (70%) than low span subjects (64%), this difference failed to yield a significant effect of span in the complex condition (F (1, 14) = 1.653). These results contradict the significant interaction between multimodality and span (a stronger speech interference effect for low than for high span subjects) and the significant main effect of span, found for complex sentences in experiment 2b.

Overall, when calculated across complexity conditions, high span subjects showed a greater resistance to speech interference than low span subjects in this experiment; this interaction between span and multimodality approached significance (F (1, 28) = 3.555; p <.08). Table 7.10 and Figure 7.12 demonstrate the relationship between these variables.

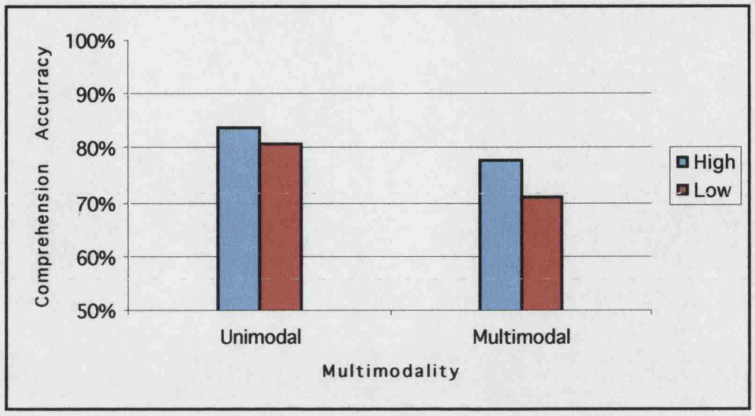
Table 7.10

Experiment 3. Mean Comprehension Rate (CR) (%): By Multimodality and Span

(Standard Errors in Parentheses)

	Multimodality		
Span	Unimodal	Multimodal	Mean difference
High	84%	78%	6%
	(2%)	(2%)	
Low	81%	71%	10%
	(2%)	(2%)	
Mean difference	3%	7%	

Figure 7.12
Experiment 3. Mean Comprehension Rate (CR) (%): By Multimodality and Span



Simple main effects showed that for low span subjects, the 10% speech interference effect is highly significant ($F(1, 14) = 34.117$; $p < .01$). For high span subjects, concurrent speech also impaired performance ($F(1, 14) = 19.261$; $p < .01$). Can these findings refute the assumed relationships between linguistic complexity and verbal WM capacity in the multimodal domain? Again, the MMUM assumed that individual differences in verbal WM capacity would affect performance when the combination of the multimodal presentation technique and the linguistic complexity of the presented materials imposes a high processing load on the users. The higher the processing load, the more apparent the difference was assumed to be. The global pattern of results does not suggest that individual differences in verbal WM capacity are more apparent as processing demands formed by the combination of the multimodal presentation technique and the linguistic complexity of the presented materials increase. For low span subjects, speech interference can be observed when processing long-simple sentences involving an immediate assignment of thematic roles. For complex sentences, the span variable fails to distinguish between varying magnitudes of multimodal interference. It is possible that more subjects (and therefore, the ability to exclude medium span subjects) are needed to validate the assumed relationships between linguistic complexity and verbal WM capacity in the multimodal domain. However, it is also possible that the assumed relationship cannot be validated using excessively complex sentences. When processing this structure, normal parse fails and processing stalls. Users must adopt a 'repair and salvage' mode in order to continue syntactic parsing, possibly requiring the parse to be restarted. The result is a non-standard processing behaviour. An intermediate level of complexity (e.g., 4Xint PLUs) might be needed to confirm these relationships.

Item analysis of the comprehension rates

Similar to the previous experiments, comprehension rates were collected for each sentence. Complexity and span formed the within-items variables while durability and multimodality formed the between-items variables. The pattern of the collected data did not warrant the use of parametric

tests[70]. However, as suggested earlier, non-parametric tests cannot control for the pragmatic complexity value of an item. Thus, in spite of the fact that the assumption of normality and of homogeneity of variance were not met as required, it was decided to perform the item analysis using parametric tests. A repeated-measure analysis of variance was therefore conducted, using the pragmatic complexity variable as a covariate.

The analysis yielded a significant effect of pragmatic complexity ($F(1, 75) = 9.696$; $p < .01$). This effect was investigated further by means of a linear regression. Consistent with the results of experiment 2b, the regression identified a significant negative relationship between pragmatic complexity and comprehension rates in this experiment: the higher the pragmatic complexity value, the lower the comprehension:
Predicted CR = (.933- .016 x Pragmatic complexity + error)
($F(1, 79) = 8.135$; $p < .01$)

Further effects replicated those found in the subjects' analysis: a significant effect of syntactic complexity ($F(1, 75) = 8.855$; $p < .01$), a significant effect of dynamism ($F(1, 75) = 4.175$; $p < .05$) and a significant effect of multimodality ($F(1, 75) = 4.509$; $p < .05$). Moreover, the interaction between complexity, span and multimodality approached significance ($F(1, 75) = 2.937$; $p < .1$). The investigation of this interaction involved a reduced analysis of variance at each level of the complexity variable. For simple sentences, the interaction between span and multimodality approached significance ($F(1, 75) = 3.426$; $p < .08$); similar to the results of the subjects' analysis, multimodality impaired the comprehension of low span users ($F(1, 78) = 10.009$; $p < .01$) and left high span users unaffected ($F(1, 78) = .230$). For complex sentences, the interaction between span and multimodality did not reach significance ($F(1, 75) = .082$); all users were affected by the added speech ($F(1, 75) = 4.715$; $p < .05$).

7.4.2 Response times

The time it took to comprehend each sentence was collected for each subject. Response time values were analysed in a repeated measure analysis of variance[71]. Complexity and span formed the

---

[70] The exploration of the items' comprehension rate data revealed that the assumption of normality was not met for all sub-conditions created by dynamism, complexity and span. In addition, the assumption of homogeneity of variance was not met for both high and low span subjects in the simple conditions. The assumption of normality was not met either for all sub-conditions created by multimodality, complexity and span. For these conditions, the assumption of homogeneity of variance was met only for high span subjects in the simple conditions. When squared values were used, they failed again to meet the assumption of normality. On the other hand, the assumption of homogeneity of variance was met for most sub-conditions. It was decided to analyse the squared values of the comprehension rates rather than the original values. Note that the reported mean values were converted back to the original units using a square root transformation

[71] An exploration of the response time data revealed that the sub-conditions created by complexity, dynamism and multimodality did not exhibit perfectly normal distributions. The assumption of normality was not met for all conditions bar the dynamic-durable simple condition (Shapiro-Wilk (16) = .875; $p < .05$). In addition, the

between-subjects variables while dynamism and multimodality formed the within-subjects variables. Response times for simple sentences (1809 ms) were significantly faster than those obtained for complex sentences (2553 ms) ($F(1, 28) = 13.751$; $p < .01$). Also, response times of high span subjects (1975 ms) were significantly faster than those of low span subjects (2387 ms) ($F(1, 28) = 5.060$; $p < .04$). These effects enhance the corresponding effects found for the comprehension rate measure but cannot shed light on the complex findings described above. No other effect reached significance.

7.5 Conclusions

In this study, two forms of multimodal presentation were used: a static-durable multimodal presentation in which processing of the visual text was user-paced and a dynamic-durable multimodal presentation in which processing of the visual (and the auditory) words was machine-paced. The use of these two multimodal techniques enabled the examination of the effect of visual-processing control in processing multimodal information given variations in syntactic complexity and thereby, also to decide between the extended durability account and the methodological account of the results of experiment 1. Results of various analyses validated some of the core assumptions made by the MMUM while refuting others.

Consistent with the assumptions made by the MMUM, multimodality was found to impair sentence comprehension under high processing demands. This finding supports the central claim of the model that the SAS relies upon the same limited pool of resources used for sentence processing for its multimodal supervision functions. When processing load increases, either due to syntactic complexity or (for low span subjects) due to sentence length, then fewer resources are available to the SAS. Consequently, its ability to supervise coordination of processing between modalities is impaired. This multimodal interference was not affected by the dynamism of the visual text, implying that the multimodal facilitation in comprehending complex sentences that was found in experiment 1 resulted from methodological limitations in the design of that dual-task experiment. The extended durability account can therefore be rejected.

---

assumption of homogeneity of variance was not met for the dynamic-durable multimodal condition, the static-durable visual condition and the static-durable multimodal condition. However, all distributions were positively skewed. Similarly, the sub-conditions created by span, dynamism and multimodality did not exhibit perfectly normal distributions. Specifically, the assumption of normality was not met for high span subjects in the dynamic-durable visual condition (Shapiro-Wilk (16) = .862; $p < .03$), in the static-durable visual condition (Shapiro-Wilk (16) = .846; $p < .02$), and in the static-durable multimodal condition (Shapiro-Wilk (16) = .819; $p < .01$). Finally, all distributions were positively skewed. An additional exploration used the natural log values of the response time data. Whereas this transformation did not affect the normality values, the assumption of homogeneity of variance was now met for all presentation conditions created by both complexity, dynamism and multimodality and span, dynamism and multimodality. It was decided to conduct the analyses of variance over the transformed values of the response time measure rather than the original values. Note that the reported mean values were converted back to the original units using an e^x transformation.

On the other hand, the study refuted the assumption that the control over visual processing assists performance only for sentences that involve an immediate assignment of thematic roles between sentential components. Regardless of syntactic complexity, multimodality and verbal WM capacity, a high visual-processing control was found to reduce user cost. Presenting the visual text in a static-durable form enables users to scan and skim the visual text and make regressive eye-movements to previously processed portions of text. Moreover, under slow presentation rate conditions, this flexible presentation form facilitates the repair process of a syntactic processing breakdown, enabling a better assignment of thematic roles between sentential constituents.

Moreover, a high visual-processing control was found to be more important than the recoverability of synchronous processing in alleviating processing cost under increasing load conditions. The MMUM claimed that, following the use of regressive eye-movements, synchronous processing would be easier to restore in the dynamic-durable multimodal format than in the static-durable multimodal format through refocusing attention on the "leading edge" of the visual display. Thus, for complex sentences, subjects were expected to experience greater speech interference in the static-durable than in the dynamic-durable conditions. In contrast, the multimodal interference effect was not affected by the dynamism of the visual text. It is therefore suggested that the importance of coupled (synchronous) presentation in recovering from an asynchronous processing was overestimated by the MMUM. Presenting the visual and the auditory words together at the same rate does not reduce the multimodal interference effect for complex sentence materials.

Finally, results did not confirm the assumed relationships between verbal WM capacity, syntactic complexity and multimodality. Individual differences in verbal WM capacity were expected to affect performance when the combination of the multimodal presentation technique and the linguistic complexity of the presented materials imposed a high processing load on the users. The higher the processing load, the more apparent the difference was assumed to be. This assumption gained only partial support. The sentence length used in both experiments 2 and 3 was capable of distinguishing between the comprehension rates of users with low and high verbal WM capacity for simple sentences. The added speech impaired performance of low span subjects (in the static-durable and the dynamic-durable simple conditions of this study and in the dynamic-transient simple conditions in the combined analysis of experiments 2a and 2b) and left high span subjects unaffected. However, in both experiments 2b and 3, the span variable failed entirely to statistically distinguish between the varying magnitudes of multimodal interference for complex sentences. Moreover, neither the durability of the visual text in experiment 2b, nor the dynamism of the visual text in this experiment was found to mediate this pattern of results. The limited number of span subjects in each complexity condition and the poor screening method of span users might suggest that abandoning the assumed relationships between linguistic complexity and verbal WM capacity in the multimodal domain is premature. Moreover, as suggested earlier, it is also possible that the assumed relationship cannot be validated using excessively complex sentences for which all subjects experienced non-normative processing (and for which a floor effect was found in the dynamic-durable multimodal condition). An intermediate level of complexity (e.g., 4Xint PLUs) might be needed to confirm these relationships.

In spite of these limitations, the varying processing capabilities of low and high span users that were identified for long-simple sentences have serious implications for effective multimodal interface design. They imply that long-simple sentences would be better presented without additional speech, if one aims to accommodate the processing requirements of all users. In conclusion, it is suggested that any guideline for multimodal presentation of sentence materials should comply with the limitations of low span users, as these are more restrictive.

This empirical work enabled the validation of some of the core assumptions made by the MMUM while refuting others. In order to improve the ecological validity of the experimental findings, there is a need to test them in less sterile situations that are more realistic in reflecting everyday settings. In the next chapter, these findings are translated to a set of guidelines for effective multimodal presentation of sentence materials. Furthermore, a final study is presented, aiming to examine the validity of some of the guidelines in an applied setting.

Chapter 8

The formulation and validation of guidelines for effective multimodal interface design

In the previous chapters, the very general design problem of when and how speech can be combined with visual text to benefit users in their work was systematically investigated, both theoretically and empirically. At the theoretical front, the research made explicit both the central features of multimodal sentence presentation (by means of the multimodal design space) and the critical structures and processes involved in multimodal language processing (by means of the MMUM). At the empirical front, the studies both investigated the central propositions of the MMUM and provided specific contents for this model through effects found.

In the first part of this chapter, the experimental findings are translated to a set of guidelines for effective multimodal presentation of sentences. In the formulation of these guidelines, the communication goals of the presented sentences are considered to be irrelevant. The guidelines are intended to be equally applicable to the presentation of conceptual information (e.g., email messages delivered on a PDA, educational materials in tutorials and leisure-type materials in entertainment systems) and procedural information (e.g., instructions for locating information in an information kiosk or for making a transaction on the internet). Hence the focus here is on the linguistic aspects of the sentences, the memory demands of the primary presentation channel and the characteristics of the user. The guidelines account for the linguistic complexity of the presented sentences and are given in terms of speech added to visual text displays for systems that are predominately visual and in terms of visual text added to speech output for systems that are predominately auditory. This distinction enables the assessment of increases and decreases in processing cost for different contents when multimodal extension is used. All guidelines aim to accommodate individual differences in verbal WM capacity so that they support the varying processing capabilities of both low and high span users that were identified in the previous studies. For example, it was found that regardless of the durability or the dynamism of the visual text, when sentences impose medium processing-load, only low span users experience multimodal interference. High span users are resistant to this interference effect under the same load conditions. These findings will be of major significance for the design of multimodal applications intended for the elderly since, as noted earlier, Just and Carpenter (1992) suggest that reduction in WM capacity occurs with ageing. Furthermore, they will be of relevance for the design of multimodal applications intended for the general population, since all types of users should be accounted for. The guidelines were therefore formed to comply with the limitations of low

span users, as these are more restrictive. A detailed explanation concerning verbal WM capacity can be found in the rationale of each guideline.

The second part of this chapter presents the final study, aiming to examine the validity of some of the guidelines in an applied setting. This setting simulates a hand held, fully mobile email system that displays static-durable messages on a palm sized screen, reads them out to the user or combines the two modes of presentation by means of a non-coupled static-durable multimodal presentation. A realistic context is assumed for the email messages to be tested and, to make them as natural as possible, only simple sentences are used. The selection of guidelines for further validation is motivated by these selections: by the scenario that the study takes, by the application type, by the display size and by the linguistic complexity of the sentences used.

8.1 Guidelines for multimodal interface design

8.1.1 Visual display devices: adding speech to visual text

The first set of guidelines refers to systems in which the visual text is the primary channel. These include extremely limited display devices (such as pagers) that cannot display a whole sentence at a time[72] (GL1) and medium to large display devices (such as medium-size hand held devices and personal computers) that present whole sentences in both dynamic and static forms (GL2). The guidelines refer to short-simple (GL1.1, GL2.1), long-simple (GL1.2, GL2.2), and long-complex sentences (GL1.3, GL2.3).

In the application of these guidelines, note that if presentation involves both short and long sentences, then consistency with long sentences should be the overriding consideration.

GL 1.1 Extremely limited display devices (e.g., pagers) – short-simple sentences

When display is extremely limited in size, present a short-simple sentence one word at a time on the centre of the screen[73].

---

[72] Extremely limited display devices present words one at a time (or a few at a time) at a fixed location on the screen (RSVP) or by means of travelling text (Times Square Format), where text is scrolled horizontally from right to left on one line displays. Given that the display window is small, these dynamic-transient presentation formats impose similar memory demands on the users. The decision of whether or not to add speech to such systems should be equal in both cases. The guidelines apply however only to RSVP of isolated words. Adding speech to the visual text in RSVP systems of chunks of words or in travelling text systems results in non-coupled dynamic presentations, neither of which was assessed in the previous studies.

[73] Note that no direct comparison was conducted between dynamic-transient visual and auditory presentation forms in the previous studies. It is assumed that both of these modes are fairly equal in terms of the processing load they impose on the user. Deciding which modality should consist of the primary presentation channel should be determined by the task context and is not within the scope of this work.

Redundant speech can be added to the visual words, at the same rate that the visual words are presented.

Rationale: Presenting visual and auditory words together at the same rate optimises lexical access of individual words. According to the MMUM, when processing demands are low, this cross-modal activation should optimise information processing by higher linguistic systems. All users are expected to have enough capacity to accommodate the storage and processing demands of short-simple sentences and the coordination of processing between modalities. Multimodality is not expected to improve comprehension due to ceiling effects. However, it might reduce lexical access times and increase overall user satisfaction.

Empirical Justification: Cross-modal priming studies, conducted for individual words.

Caveat: This configuration was not explicitly tested.

GL 1.2 Extremely limited display devices (e.g., pagers) – long-simple sentences

When display is extremely limited in size, present a long-simple sentence one word at a time on the centre of the screen.
Avoid adding redundant speech to the visual words.
Allow users to replay the sequence.

Rationale: Low span users do not have enough capacity to accommodate both storage and processing demands of long sentences. For these users, concurrent speech interferes with the mental retrieval of previously processed information. In contrast, high span users have enough capacity to accommodate the storage and processing demands of long-simple sentences and the coordination of processing between modalities. Concurrent speech does not interfere with the mental retrieval of previously processed information for these users.
Allowing users to replay the transient sequence will facilitate its comprehension.

Empirical justification: See combined analysis of experiments 2a and 2b, comprehension rates of low and high span subjects for long-simple sentences in the transient conditions.

GL 1.3 Extremely limited display devices (e.g., pagers) – long-complex sentences

Obviously, these devices are not suitable for the presentation of long-complex sentences.
When unavoidable, present the visual words one at a time on the centre of the screen.
Do not add speech to the visual display under any circumstances.

Rationale: All users do not have enough capacity to accommodate both storage and processing demands of long-complex sentences. Speech must be avoided; not only is the simultaneous multimodal activation of the cross-modal systems negligible under a high processing load but the

delayed assignment of thematic roles necessary for comprehension of complex sentences is impaired by the concurrent speech.

Empirical justification: See experiments 2a and 2b; comprehension rates of all subjects for long-complex sentences in the transient conditions.

GL 2.1 Medium to large display devices (e.g., medium-size hand held devices, personal computers) – short-simple sentences

When display size does not limit the mode of presentation of a short-simple sentence, it is preferable to use a static-durable form rather than a dynamic-durable form.
Speech can be added to the static-durable visual text.

Rationale: A static-durable presentation enables faster visual processing of the sentence than a dynamic-durable presentation in which reading pace is machine-governed. The choice of visual format should not affect comprehension rate. Similarly, multimodality is not expected to affect comprehension rate. Under these low processing demands, all users are expected to have enough capacity to accommodate the storage and processing demands of short-simple sentences and the coordination of processing between modalities. However, multimodality might reduce lexical access times and increase overall user satisfaction.

Empirical justification: See experiment 1, word-monitoring data for early target words in the visual-based conditions.

Caveat: This configuration was not explicitly tested.

GL 2.2 Medium to large display devices (e.g., medium-size hand held devices, personal computers) – long-simple sentences

When display size does not limit the mode of presentation of a long-simple sentence, use a static-durable form rather than a dynamic-durable form. This is especially important when the display of the sentence is spread over more than one line.
Speech should not be added to the visual display.

Rationale: Again, a static-durable visual presentation enables faster and more flexible visual processing of the sentence than a dynamic-durable presentation in which reading pace is machine-governed, and therefore reduces differences between low and high span users. Furthermore, in cases where display of the sentence is spread over more than one line, the dynamic-durable visual presentation impairs users' ability to predict the location in which words are to appear (users cannot predict whether a short word will appear at the end of the same line, or a longer word at the beginning of the next line).

Speech should be avoided; when low span users process long-simple sentences, concurrent speech interferes with the recollection of intermediate representations. High span users are resistant to such an interference effect. These users have enough capacity to accommodate the storage and processing demands of long-simple sentences and the coordination of processing between modalities.

Empirical justification: For the effect of number of lines in a dynamic-durable presentation, see item analysis in experiment 2a; poorer comprehension rates of double-line sentences relative to single-line sentences. For the advantage of static over dynamic-durable presentation, see experiment 3; higher comprehension rates in the static conditions than in the dynamic conditions found for simple sentences in both the unimodal and the multimodal conditions. Also in experiment 3, see speech interference effects in comprehending long-simple sentences by low span subjects.

GL 2.3 Medium to large display devices (e.g., medium-size hand held devices, personal computers) – long-complex sentences

Presenting a long-complex sentence is not advisable.
When unavoidable, a static-durable visual presentation is the best form of presentation.
Do not add speech to the visual display under any circumstances.

Rationale: All users do not have enough capacity to accommodate both storage and processing demands of long-complex sentences. However, a static-durable visual presentation should optimise the repair process of the syntactic processing breakdown.
Speech must be avoided; the recollection of intermediate representations using regressive eye-movements is greatly reduced by concurrent speech.

Empirical justification: For the effect of visual processing control in comprehending complex sentences, see experiment 3; higher comprehension rates in the static conditions than in the dynamic conditions found for complex sentences in both the unimodal and the multimodal conditions.
Also in experiment 3, see speech interference in comprehending complex sentences.

8.1.2 Speech-based systems: adding visual text to speech

The second set of guidelines refers to systems in which speech is the primary channel (e.g., in PA systems and telephone-based services). As a dynamic media, speech attracts user attention automatically. Furthermore, due to its transient nature, speech must be processed as it arrives. It is assumed that a static-durable visual text is the best option to backup the auditory message in such systems. For example, flight connection information read out over an in-flight sound system while being echoed by a static visual display of text on the airplane's video system. Hence, the following guidelines do not refer to a dynamic-transient, or to a dynamic-durable visual backup.

Again, note that if presentation involves both short and long sentences, then consistency with long sentences should be the overriding consideration.

GL 3.1 Speech-based systems - short-simple sentences

A static-durable text can be added to a short-simple spoken sentence.

Rationale: All users are expected to have enough capacity to accommodate both storage and processing demands of short-simple spoken sentences. The addition of a static-durable visual text is not expected to affect comprehension. However, it might reduce lexical access times and increase overall user satisfaction.

Caveat: This configuration was not explicitly tested.

GL 3.2 Speech-based systems - long-simple sentences

A static-durable visual text should be used to echo the primary auditory channel if the message is relatively long.

Rationale: Since speech must be processed as it arrives, displaying the message in a durable visual form would ease a further retrieval of intermediate representations by all users, especially in a noisy environment. This redundant presentation would be inferior to a static-durable presentation of the visual message because for low span users, concurrent speech interferes with the recollection of intermediate representations. High span users are resistant to this interference effect. These users have enough capacity to accommodate the storage and processing demands of long-simple sentences and the coordination of processing between modalities.

Empirical justification: For the facilitation of processing by static-durable visual text, see experiment 1; comprehension rates of long-simple sentences in the speech-based conditions for all subjects. For speech interference effects in comprehending long-simple sentences by low span subjects, see experiment 3; comprehension rates of low span subjects for long-simple sentences in the static-durable conditions.

GL 3.3 Speech-based systems - long-complex sentences

Again, presenting a long-complex sentence is not advisable.
A static-durable visual text should be used to echo the primary auditory channel if the message is long and complex.

Rationale: All users do not have enough capacity to accommodate both storage and processing demands of long-complex sentences. However, adding a static-durable visual text to the primary auditory channel helps the repair process of the syntactic processing breakdown. This redundant presentation is inferior to a unimodal static-durable presentation of the visual message because for all users, the recollection of intermediate representations using regressive eye-movements is greatly reduced by concurrent speech.

Empirical justification: For the facilitation of processing by static-durable visual text, see experiment 1; comprehension rates of all subjects for long-complex sentences in the speech-based conditions. For speech interference effects in comprehending long-complex sentences, see experiment 3; comprehension rates of all subjects for long-complex sentences in the multimodal relative to the unimodal conditions.

Caveat: The multimodal facilitation in comprehending complex sentences that was found in experiment 1 is confounded by the relative ease of the word-monitoring task in the multimodal conditions. Note that this is the only experiment that directly assesses differences between spoken presentation and a static-durable multimodal presentation of complex sentences.

8.2 The applied study

8.2.1 Introduction

The translation of the experimental findings enables the formation of a set of guidelines for effective presentation of sentences, essential to achieving the primary goal of informing the design process of multimodal user interfaces. However, whilst the guidelines are derived through evidence, their validity for interface design cannot simply be assumed without further testing. Inevitably, there will arise additional factors in design situations, which require further qualification of the guidelines (for example, the effect of situating sentences within paragraphs). Also, the guidelines as presented may not have the same level of validity for a number of reasons:

Minimal empirical evidence: The guidelines constructed for short-simple sentences were based on minimal empirical evidence, namely:

- Cross-modal priming studies conducted for individual words reported in the literature.
- Word-monitoring data of early target words in the visual-based conditions that was observed in experiment 1.
- Speech interference continuum found for varying processing load levels, where load is a function of both linguistic complexity and individual verbal WM capacity. Specifically, it was found that when sentences impose an excessive load, all users experience speech interference. On the other hand, when sentences impose a medium load, only low span users experience speech interference. Assuming that short-simple sentences impose a low processing load on all users, no speech interference should be apparent; all users are expected to have enough capacity to accommodate the storage and processing demands of short-simple sentences and the coordination of processing between modalities.

None of the investigative studies (experiments 1 to 3) actually manipulated sentence length; it is yet to be determined whether multimodality reduces processing cost of short-simple sentences (by means of reducing lexical access times and increasing the overall satisfaction of the user).

Low ecological validity: The studies aimed to provide coherent explanations of multimodal language processing, therefore constraints of a laboratory setting were devised to maximise experimental control. This is why only two sentence structures were sampled, a single sentence length was used, context was eliminated and intonation and prosody were kept to a minimum. In order to improve the ecological validity of the guidelines, they need to be tested in less sterile situations that reflect real everyday settings more closely.

An applied setting was defined to test the validity of some of these guidelines. This setting is based on the projection of a multimodal multimedia device that was recently attempted by BT Cellnet (see Chapter 1). The setting simulates a hand held, fully mobile email system that can display static-durable messages on a palm sized screen, present them via earphones/speaker or combine the two by means of a static-durable multimodal presentation. Other characteristics of the device are not represented in this setting (e.g., speech recognition features, presenting pictures and movies). In spite of its basic characteristics, the simulation enables examination of a cross section of the guidelines, though not all of them:

- It is a good platform for testing the guidelines for medium to large display devices as the screen size enables presentation of full-length sentences. Guidelines for medium to large display devices can be assessed by means of comparing the static-durable visual format with the non-coupled static-durable multimodal format. The platform may also be used for testing guidelines relating to the dynamic-transient presentation format, but since this is machine-paced it is assumed to be less effective.

- It is a good platform for testing the guidelines for speech-based systems. In this simulation, email messages can be read out to the user, freeing their hands and vision while display size enables backing up the spoken message by means of a static-durable visual text. Guidelines for speech-based systems can be assessed by means of comparing the auditory format with the non-coupled static-durable multimodal format.

- It is a good platform for testing natural sentences in "real life" scenarios.
  In this study, a realistic context is assumed for the email messages to be tested. For example, in a typical scenario, subjects are asked to imagine that they manage a designers' recruitment agency. The responsibilities of the job (e.g., recruiting designers for various jobs and projects, management of employees, etc.) are listed to provide context for the natural email messages that follow.
  Appendix D presents the email messages used in this study. All email messages used are a mix of short-simple priming sentences followed by either short (2 clauses) or long (3 clauses) right-branching test sentences, for example[74]:

[74] Different sentences were used in this study. The same "core" sentence is presented here for demonstration only.

Priming sentence:

The country inn project is in hand.

Short right-branching sentence:

Suzan Archer interviewed Angela Roberts who is currently working for Future Interiors.

Long right-branching sentence:

Suzan Archer interviewed Angela Roberts who is currently working for Future Interiors that designed the Four Seasons hotel.

Guidelines for short and long-simple sentences can be assessed by means of comparing subjects' comprehension rates of short and long right-branching sentences. In addition, although the study does not capture lower level factors such as lexical access times, it does examine the possibility that for low processing load levels, multimodality improves the overall satisfaction of all users relative to each of the unimodal conditions. Appendix E presents a preference questionnaire used in this study. Users' preferences are presented in Appendix F and are referred to in the results' section of this chapter. This natural approach is not suitable for testing the guidelines of excessively complex sentences.

- The applied setting also allows for testing predicted differences between low and high span users specified in the rationale of each guideline.

8.2.2 Experimental hypotheses

This section provides detailed predictions for the comprehension rate (CR) measure in this experiment. The predictions do not refer to the time it takes subjects to respond to each comprehension statement as, based on the results of the investigative studies, it was assumed that none of the experimental factors would affect the time it takes to comprehend each statement. Response times were simply collected as a measure of control against unexpected trade-offs between the speed of responses and their accuracy.

Table 8.1 summarises the predictions of the guidelines to be tested by the applied setting. It distinguishes between the predictions made for medium to large display devices (GL2) and those made for speech-based systems (GL3). Another distinction is made between the predictions made for short and long sentences in each set of guidelines (GL2.1 & 3.1 and GL2.2 & 3.2 respectively). Further sub-predictions differentiate between the expected CR of high and low span users in each presentation condition for each level of sentence length in this study. The predictions are systematically reviewed in the discussion of the results of the study in section 8.2.4.

Table 8.1

Applied Study. Predicted Comprehension Rates (CR) made by the Guidelines

Prediction made by the guidelines		confirmation of the prediction by the applied study
GL 2	**(medium to large display devices)**	
GL 2.1 short	$(CR_{HS\,V} = CR_{LS\,V}) = (CR_{HS\,MM} = CR_{LS\,MM})$	
P 2.1.1 short	$(CR_{V} = CR_{MM})$	√
P 2.1.2 short	$(CR_{HS} = CR_{LS})$	X
P 2.1.3 short	$(CR_{HS\,V} - CR_{LS\,V}) = (CR_{HS\,MM} - CR_{LS\,MM})$	√
GL 2.2 long	$(CR_{HS\,V} = CR_{HS\,MM} \cong CR_{LS\,V}) > CR_{LS\,MM}$	
P 2.2.1 long	$(CR_{V} > CR_{MM})$	X
P 2.2.2 long	$(CR_{HS} > CR_{LS})$	√
P 2.2.3 long	$(CR_{HS\,V} - CR_{LS\,V}) < (CR_{HS\,MM} - CR_{LS\,MM})$	X
GL 3	**(speech-based systems)**	
GL 3.1 short	$(CR_{HS\,A} = CR_{LS\,A}) = (CR_{HS\,MM} = CR_{LS\,MM})$	
P 3.1.1 short	$(CR_{A} = CR_{MM})$	√
P 3.1.2 short	$(CR_{HS} = CR_{LS})$	X
P 3.1.3 short	$(CR_{HS\,A} - CR_{LS\,A}) = (CR_{HS\,MM} - CR_{LS\,MM})$	√
GL 3.2 long	$(CR_{HS\,MM} > CR_{HS\,A}) > (CR_{LS\,MM} > CR_{LS\,A})$	
P 3.2.1 long	$(CR_{A} < CR_{MM})$	√
P 3.2.2 long	$(CR_{HS} > CR_{LS})$	√
P 3.2.3 long	$(CR_{HS\,A} - CR_{LS\,A}) = (CR_{HS\,MM} - CR_{LS\,MM})$	√

Index:

HS - high span, LS – low span, V – visual, MM – multimodal, A – auditory,

√ - confirmed prediction, X – rejected prediction

When viewed together, the guidelines tested by this study predict a triple interaction between span, sentence length and mode of presentation. It is expected that low span users will experience an interference in the multimodal condition relative to the visual only condition to which high span users will be immune, whilst both types of user will experience a facilitation of the multimodal condition relative to the auditory only condition. The source of this triple interaction is the expected interaction between presentation and span for the long (3 clauses) right-branching sentences. The expected pattern of results for the short (2 clauses) right-branching sentences follows.

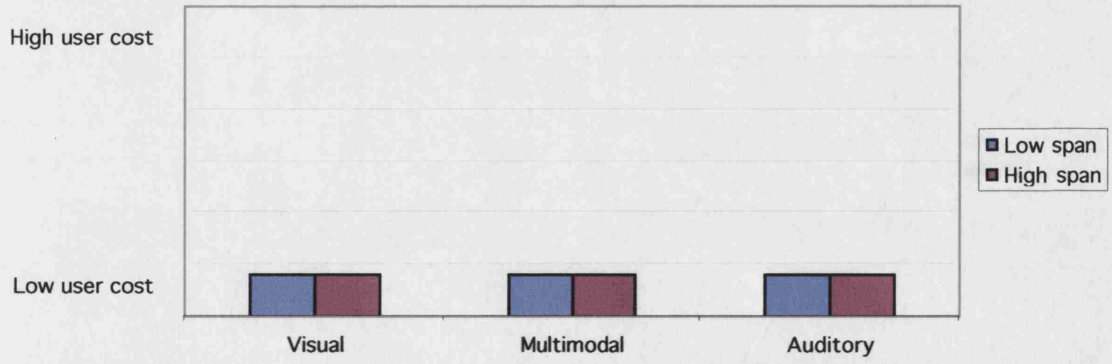
Assuming that the short right-branching sentences place a low processing load on all users, GL 2.1 and GL 3.1 predict that both high and low span subjects will show high comprehension rates of these sentences in all presentation conditions (see also Figure 8.1). Specifically, according to GL 2.1, the multimodal condition should not impair performance in relation to the static-durable visual condition: all users are expected to have enough capacity to accommodate the storage and processing demands of short-simple sentences and the coordination of processing between modalities. For these sentences, GL 3.1 predicts that multimodality will not improve comprehension rates in relation to the auditory condition: all users are expected to have enough capacity to accommodate both storage and processing demands of short-simple sentences, and the high level of visual processing control in the multimodal condition will not assist a further retrieval of intermediate representations.

In summary, for the short sentences, GL 2.1 and GL 3.1 predict:

- No effect of presentation (P 2.1.1 and P 3.1.1).
- No effect of span (P 2.1.2 and P 3.1.2).
- No interaction between span and presentation (P 2.1.3 and P 3.1.3).

Figure 8.1

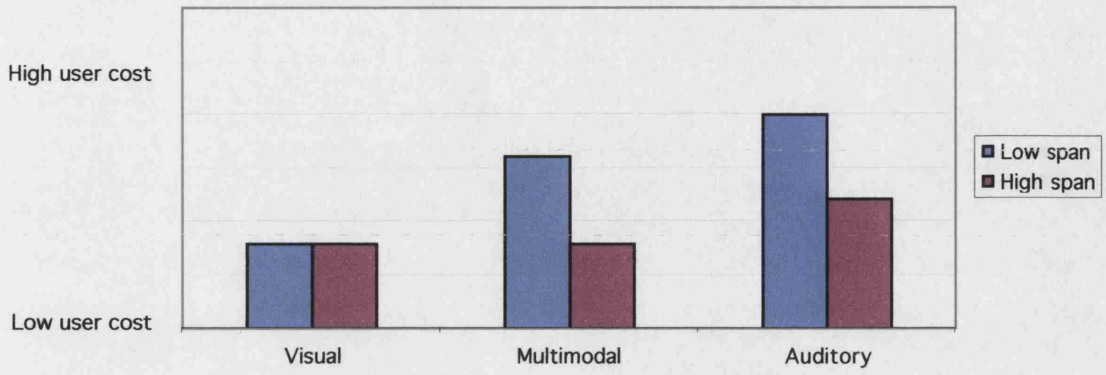
Predicted User Cost as a function of Presentation Type and Individual Verbal WM Capacity for 2 Clauses (Short) Right-Branching Sentences (X_{INT} PLUs)



The long right-branching sentences are assumed to impose a higher processing load on all users and consistent with the previous studies, to show differences between comprehension rates of low and high span users. When processing load increases, users may rely on the durability of the visual text using regressive eye movements. According to GL 2.2, this reliance should diminish any differences between low and high span subjects in the visual condition (see Figure 8.2).

Figure 8.2

Predicted User Cost as a function of Presentation Type and Individual Verbal WM Capacity for 3 Clauses (Long) Right-Branching Sentences (X_{INT} PLUs)



The same guideline predicts however differences between span subjects in the non-coupled static-durable multimodal condition. In this condition, regressive eye-movements are conducted while attending to the spoken continuation of the sentence. As long as the available resources are sufficient, processing may proceed with no interference. This is precisely the pattern of behaviour predicted for high span subjects at this level of load (see Figure 8.2). On the other hand, when resources are low, the recollection of the information during the visual regression will be impaired by concurrent speech due to the failure of the SAS to supervise coordination of processing of visual and auditory information. Figure 8.2 demonstrates this predicted state of affairs for low span subjects; these subjects will not have enough capacity to accommodate the storage and processing demands of long right-branching sentences and the coordination of processing between modalities. They will therefore experience speech interference at this level of load.

GL 3.2 predicts that this multimodal interference experienced by low span subjects should turn into multimodal facilitation when compared with the dynamic-transient auditory condition. Since speech must be processed as it arrives, the durable backup provided by the visual text may ease a further retrieval of intermediate representations. According to the same guideline, high span subjects are also expected to benefit from the high level of visual processing control in the multimodal condition (see Figure 8.2).

In summary, for the long sentences, GL 2.2 and GL 3.2 predict:

- An effect of presentation: a multimodal interference effect relative to the visual condition and a multimodal facilitation effect relative to the auditory condition (P 2.2.1 and P 3.2.1).
- An effect of span: higher comprehension rates of high span subjects relative to low span subjects (P 2.2.2 and P 3.2.2).
- An interaction between span and presentation. Specifically, whereas for the visual contrast,

P 2.2.3 predicts a strong multimodal interference effect for low span subjects and no effect for high span subjects, P 3.2.3 predicts a multimodal facilitation effect for both high and low span users relative to the auditory condition.

Averaged over sentence length, a significant interaction between presentation and span is expected to be found: comprehension rates of high span users will be equally high regardless of the presentation format. In contrast, comprehension rates of low span users will be affected by the different modes of presentation. The source of this expected interaction is their predicted comprehension of long right-branching sentences in the different presentation conditions, as described above.

The absence of such differences, predicted for the short right-branching sentences, will produce a significant interaction between length and span. In addition, since comprehension rates of high and low span users are expected to differ only in the auditory and the multimodal conditions of long sentences, a main effect of span is not predicted in this study.

Moreover, regardless of users' verbal WM capacity, a significant interaction between length and presentation is predicted: whereas comprehension rates of short sentences will be equally high regardless of presentation, those of long sentences will vary depending on the different presentation conditions. For these sentences, a multimodal interference effect relative to the visual condition and a multimodal facilitation effect relative to the auditory condition are expected. On average, a main effect of presentation is predicted. The source of this effect should be the difference between the visual and the auditory conditions for the long right-branching sentences; the predicted differences between the multimodal condition and each of the unimodal conditions are not expected to be large enough to contribute to this main effect. Finally, the overall differences between short and long right-branching sentences are expected to be large enough to produce a main effect of length in this study.

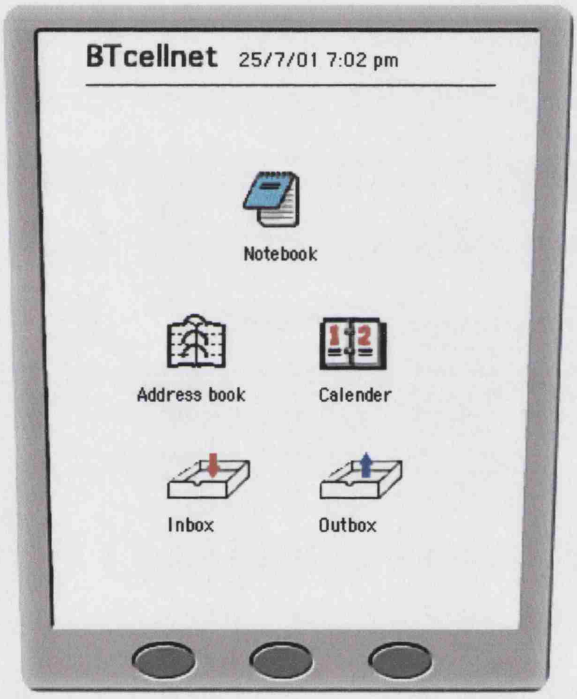
8.2.3 Method

Materials and design

The applied study was conducted using Director™ version 7.0 and was run on a laptop.

The system attempted to simulate a real hand held organiser. The header, permanent through all the screens, included the brand logo, the date and the time. Another common element was the drawing of an abstract frame (4.5 by 3.75 inch) with three inactive buttons. Figure 8.3 presents the first introductory screen.

Figure 8.3 Applied Study. The Introductory Screen



This screen included 5 labelled icons, of which only the Inbox icon was active. Clicking on this icon lead the user to the Inbox screen. The Inbox screen included a number of active components: a list of email titles, a "new messages" counter and a scroll bar. The list of messages included 19 email titles. Bar the first title (a dummy trial), the order of all other titles was randomised for each subject. Figure 8.4 shows that only 15 titles could be seen at once on the screen. The user could scroll up and down the screen by clicking the arrows or dragging the scroll handle to see messages that were not in current view.

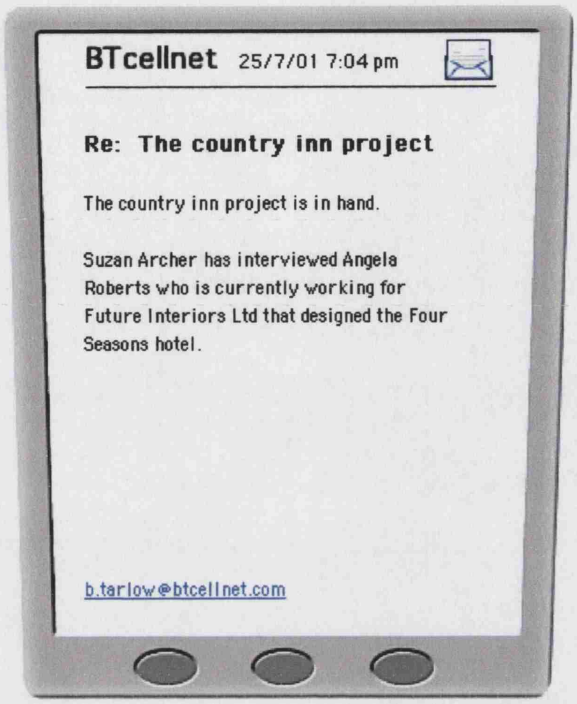
Figure 8.4 Applied Study. The Inbox Screen

The experiment included three separate presentation conditions: a static-durable visual text presentation, a static-durable multimodal presentation and an auditory presentation[75]. In all conditions, including the auditory condition, the selection of messages was by pointing and clicking. Clicking on a message title or on its envelope icon opened the selected message. Back in the Inbox, titles of opened messages changed colour to gray and their adjacent icons changed both shape and colour (from closed-yellow to opened-gray envelopes). Once selected, these titles could not be selected again. Finally, progress within the Inbox screen was indicated by the active counter that showed a running count of unopened messages.

Each email message consisted of five components (see Figure 8.5): the header (including the open envelope icon), the message title, the first (priming) sentence, the second (test) sentence and the sender's details.

[75] See Chapter 6, footnote 33

Figure 8.5 Applied Study. An Email Screen



The header, the sender's details and the message title were presented from the onset of the screen. After 2 seconds, the priming sentence was presented – visually and/or aurally – for a duration equal to the time it took to say that sentence (even in the visual only condition). Each priming sentence was devised so as to put its test sentence in context (note that each message title was formed from the priming sentence to minimise its effect on the test sentence). Priming sentences were selected by three judges; they were required to be as simple as possible and to be easily linked to the test sentences. Since the strength of that effect varies from one sentence to another, 11 judges were asked to rate how easy it was to form a connection between the priming and the test sentences in all presentation conditions. This was to ensure that no major differences exist between conditions. The scores are reported in the results' section of this chapter.

The test sentence (right-branching structure) was the final addition to the screen and, like the priming sentence, it was displayed for a period equivalent to its spoken duration (the priming sentence remained on the screen in the visual-based conditions). Each subject was presented with short and long test sentences in each of the presentation conditions. Of the 19 trials in each presentation condition, 8 trials consisted of short test sentences (between 11 and 16 words each), 8 trials consisted of long test sentences (between 17 and 21 words each), and 3 trials were dummy and filler trials. Test sentences were balanced across presentation conditions with respect to their length (on average: 15.7

words) and their pragmatic complexity value (7.92 for the short test sentences and 12.83 for the long test sentences - on average: 10.38)[76].

In each email message in all presentation conditions, the header, the sender's details and the message title were presented by static-durable visual text. Presentation of the priming and the test sentences differed between conditions. In both the static-durable visual-only and the multimodal condition, presentation of the priming and test sentences was made using Geneva 10 points font. Digital recording and editing of the priming and test sentences in the speech and multimodal conditions was made using the SoundEdit™ application (sample rate: 22.05 kHz, sample size: 16 bits). The sentences were spoken in a female voice, this time with slight intonation and prosody and an average presentation rate of 153 words per minute. Presentation duration of each sentence took on average 5957 ms.

A different scenario accompanied each of the three presentation conditions to provide a general context for the email messages. The scenarios do not directly affect the analysis of the test sentences[77]:

- Static-durable visual text condition (Designers Recruitment scenario): "You are the manager of a designers recruitment agency. Your responsibilities include recruiting designers for various jobs and projects, management of your employees, etc."
- Auditory condition (Film scenario): "You are the producer of a new film. Your responsibilities include recruiting creative and technical people, management of the people on the set, etc."
- Non-coupled static-durable multimodal presentation condition (Law Firm scenario): "You are the manager of a law firm. Your responsibilities include legal issues, recruitment issues, management of your employees etc."

---

[76] See Chapter 5, footnote 25.

[77] Applying a more complete context to the scenarios was considered: in this study, each of the test sentences consists of two names for which the subjects did not receive any prior information. In real life, the reader usually has a stronger contextual knowledge when processing such a message; at least one person related to the task at hand, their roles and responsibilities are known in advance. Subjects could be briefed regarding to who these people were, how they relate to one another and what they do in the organisation. However, due to the number of names, this was found to be unrealistic. Also, becoming familiar with names of people related to the task at hand involves creating elaborate mental models of the situation. This introduces the risk that subjects who are competent in creating models will do better than those who are less competent. Most importantly, supplying a strong contextual knowledge would likely reduce the effects sought with regard to verbal WM capacity. Specifically, a strong contextual knowledge might make a long right-branching sentence as easy to process as a short right-branching sentence without such knowledge. Consider again the long right-branching example: *Suzan Archer has interviewed Angela Roberts who is currently working for Future Interiors Ltd that designed the Four Seasons hotel.* If subjects are told that *Suzan Archer* works in the recruitment agency, it is less probable that low span subjects would make a mistake in answering that *Suzan Archer is currently working for Future Interiors*, regardless of the format of the sentence presentation. Such a strong contextual knowledge could easily mask the phenomena in question.

Comprehension statements

A "Yes" or "No" comprehension statement was created for each test sentence (including the dummy and filler trials), in order to assess whether subjects were properly processing each presented sentence. To compare the comprehension of 2 clauses with that of 3 clauses sentences, all comprehension items were constructed by combining one of the first two verbs with two of the first three nouns. (e.g., *Suzan Archer has interviewed Angela Roberts who is currently working for Future Interiors Ltd that designed the Four Seasons hotel;* statement: *Has Suzan Archer interviewed Angela Roberts? - Yes*). Comprehension statements were balanced across presentation conditions with respect to the number of true/false statements[78] and their internal distribution of the nouns-verb combinations. Pragmatically meaningless items were excluded.

Dummy and filler trials

The list included two long right-branching filler messages. The aim of these messages was to eliminate the possibility that subjects would ignore the third clause, which would not be probed for otherwise. Comprehension questions of these messages probed for the third clause in both cases. The dummy trial was always the first message in the list and had to be opened first. This long right-branching message served the simple purpose of allowing the subjects to get used to the system at the beginning of each mode of presentation. It also probed for the third clause.

Implementation

Six experimental versions were created, each had a different order of the presentation conditions to minimise order effects[79]:
Version 1: Visual, Auditory and Multimodal
Version 2: Auditory, Multimodal and Visual
Version 3: Multimodal, Visual and Auditory
Version 4: Visual, Multimodal and Auditory
Version 5: Auditory, Visual and Multimodal
Version 6: Multimodal, Auditory and Visual

Apparatus

The experiment was run on a 300 MHz ibook. Spoken sentences were presented at a comfortable volume through headphones (Vivanco SR 250) and the visual material was presented on the ibook 14 inch display (screen resolution: 1024 X 768, 75 Hz). Subjects' responded via the ibook USB keyboard.

---

[78] See Chapter 5, footnote 26.

[79] See Chapter 5, footnote 27.

Procedure

Subjects first read a simple set of instructions, describing the system, the experimental task and the email structure they would encounter, illustrated with an example. They were also told that they would pass through three simple scenarios, each relating to an imaginary work situation in which they would be receiving emails. Each scenario would consist of a different mode of presentation of the emails: visual text, speech or a combined mode. Subjects were told that their task consists of answering comprehension questions to test their understanding of each presented message. It was explained that after the message disappeared, they would be presented with a question. Their task would be to judge whether the answer to the question is true or false by pressing the "Z" key (labelled 'Yes') or the "X" key (labelled 'No'), respectively. They were told to respond as quickly and accurately as they could and that an alert sound (a "Beep") would be given for an erroneous response. The execution of a response would immediately return them to the inbox screen where they would be expected to continue with the next email message.

After confirming that the instructions were understood, the practice session started. The practice covered the 3 different presentation conditions described above. Subjects were presented with specific instructions for each presentation condition followed by 9 practice messages (see Appendix D). Most of the comprehension questions of the long right-branching sentences in the practice session probed for the third clause. This was intended to establish in subjects' the expectation that the question would refer to the third clause and therefore, they would need to attend to it.

At the end of the experimental session, subjects were interviewed by the experimenter. Questions covered personal details of computer experience (e.g., email usage time), subjective views of the different presentation techniques they had experienced in the experiment, strategies used during the experiment, user preferences and further comments about their views on the potential value of multimodality. The questionnaire is presented in Appendix E. Users' data is summarised in Appendix F.

Span task

Subjects' verbal WM capacity was measured at the beginning of the session. This was followed by a 10 minute break. The same testing procedure was followed as in the previous experiments. 16 subjects whose reading spans were 3.5 or higher were classified as high span subjects and 16 subjects whose reading spans were 3.0 or lower were classified as low span subjects.

Subjects

32 subjects were tested and were paid £10 for their participation. English was the first language of all subjects.

8.2.4 Results

Priming sentences

11 judges rated how easy it was to form a connection between the priming and the test sentences on a scale of 1 to 10 (1 - very easy, 10 – very difficult) in all presentation conditions. The mean scores are presented in Table 8.2.

Table 8.2
Applied Study. Mean Ease of Forming the Connection between Priming and Test Sentences: By Length of Test Sentences and Presentation Conditions (Standard Errors in Parentheses)

	Length By Presentation		
	Visual	Multimodal	Auditory
Short	3.57	3.50	3.16
	(.31)	(.33)	(.44)
Long	3.32	3.43	2.93
	(.42)	(.47)	(.34)
Mean difference	.25	.007	.23

A repeated-measure analysis of variance was conducted over these rates[80]. It reveals that forming the connection between the priming and the test sentences did not differ between presentation conditions ($F(2, 6) = .797$). Also, forming the connection was not affected by the length of the test sentences ($F(1, 7) = .146$). Finally, the two factors failed to interact ($F(2, 6) = .026$). This ensures that no major differences exist between conditions that could otherwise account for the pattern of results in this study.

Comprehension rates and preference data

Comprehension rates were collected for each subject and were analysed in a repeated-measure analysis of variance. Span formed the between-subjects variable while length and mode of presentation formed the within-subjects variables[81],[82]. The results only partially supported the

[80] An exploration of the judgement rates revealed that the assumptions of normality and homogeneity of variance were fully met for all sub-conditions created by the length and mode of presentation variables.

[81] An exploration of the comprehension rate data revealed that for the sub-conditions created by the span, length and mode of presentation variables, the assumptions of normality and homogeneity of variance were not met for most sub-conditions. However, it was found that 10 out of the 12 sub-conditions were negatively skewed. This provided the justification to perform additional exploration using transformed CR values. When CR values were raised to the power 4, the assumption of homogeneity of variance was met for most sub-conditions. It was decided to conduct an analysis of variance over the transformed values of the CR measure

predictions of the guidelines. Most importantly, the results failed to confirm the assumed relations between span, sentence length and mode of presentation (F (2, 29) = .590). In brief, short right-branching sentences were assumed to place a low processing load on all users, yielding equally high comprehension rates in all presentation conditions (see GL 2.1 and GL 3.1). The long right-branching sentences were assumed to impose a higher processing load and to show differences between performance of high and low span subjects. Specifically, for high span subjects, a multimodal facilitation effect was predicted relative to the auditory condition, whereas for low span subjects an additional multimodal interference effect was predicted relative to the visual condition (see GL 2.2 and GL 3.2). The mean CR for the two span groups by the three presentation conditions in the short right-branching condition are given in Table 8.3. The relationship between these means can be seen graphically in Figure 8.6.

Table 8.3

Applied Study. Mean Comprehension Rates (CR) for Short (2 Clauses) Right-Branching Sentences (%): By Span and Presentation (Standard Errors in Parentheses)

	Span By Presentation		
	Visual	Multimodal	Auditory
High span	97%	95%	98%
(N=16)	(2%)	(3%)	(2%)
Low span	90%	87%	86%
(N=16)	(2%)	(3%)	(2%)
Mean	94%	91%	92%

Consistent with the experimental predictions of GL 2.1 and GL 3.1, comprehension rates of short right-branching sentences were not affected by the mode of presentation (F (2, 29) = .355). The multimodal condition did not impair performance in relation to the static-durable visual condition (c.f., P 2.1.1), neither did it affect comprehension rates relative to the auditory condition (c.f., P 3.1.1). Also, as predicted, results indicate that the span and the presentation variables did not interact for this level of load (F (2, 29) = 1.129); (c.f., P 2.1.3 and P 3.1.3). For high span subjects, there was no effect of presentation in the visual-multimodal contrast (F (1, 15) = 1.000), in the auditory-multimodal contrast (F (1, 15) = 1.901) or in the visual-auditory contrast (F (1, 15) = .190). Similarly, for low span subjects, there was no effect of presentation in the visual-multimodal contrast (F (1, 15) = .128), in the auditory-multimodal contrast (F (1, 15) = .584) or in the visual-auditory contrast (F (1, 15) = 1.279).

---

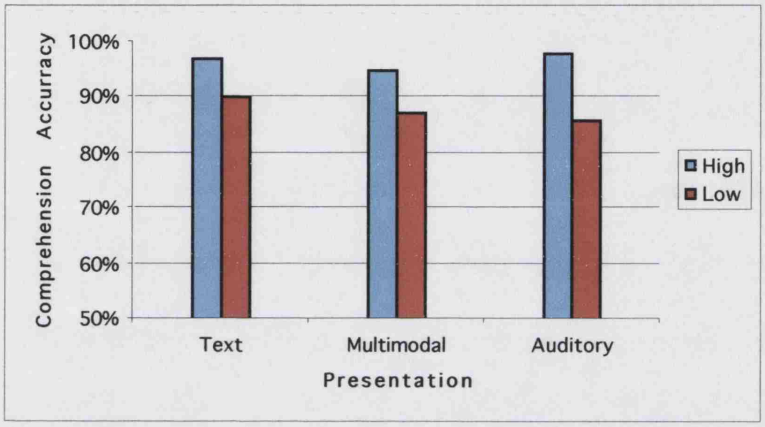
rather than the original values. Note that the reported mean values were converted back to the original units using a fourth root transformation.

[82] An additional analysis was conducted at the request of the examiners of this dissertation. This analysis included the order in which subjects performed the three presentation conditions as an additional between-subjects variable (called version number) to make sure that the experimental results were unaffected by order effects (see Method, section 8.2.3). None of the results reported in this section was affected by the order in which subjects performed the three presentation conditions. Note that the number of subjects in each version is too small to make this a reliable conclusion.

Subjects' preference data, presented in Appendix F, supports these equal comprehension rates. Regardless of their memory span, when asked which presentation best supported the shorter test sentences, users' responses distributed fairly equally between conditions; 25% of the subjects preferred the multimodal condition, 31% the visual condition, 19% the auditory condition, 16% found no difference between conditions and 9% were not sure.

Figure 8.6

Applied Study. Mean Comprehension Rates (CR) for Short (2 Clauses) Right-Branching Sentences (%): By Span and Presentation



On the other hand and in contrast to predictions P 2.1.2 and P 3.1.2, comprehension rates of high span subjects were overall 9% higher than those of low span subjects for the short right-branching sentences (F (1, 30) = 12.907; $p < .01$). Thus the 2 clauses right-branching sentences were sufficiently demanding to show differences between high and low span subjects, yet did not produce an effect of presentation mode.

The pattern of data found for the longer 3 clauses right-branching sentences is also only partially compatible with the experimental predictions.

Table 8.4

Applied Study. Mean Comprehension Rates (CR) for Long (3 Clauses) Right-Branching Sentences (%): By Span and Presentation (Standard Errors in Parentheses)

	Span By Presentation		
	Visual	Multimodal	Auditory
High span	95%	92%	86%
(N=16)	(3%)	(4%)	(3%)
Low span	86%	78%	75%
(N=16)	(3%)	(4%)	(3%)
Mean	91%	85%	81%

Figure 8.7

Applied Study. Mean Comprehension Rates (CR) for Long (3 Clauses) Right-Branching Sentences (%): By Span and Presentation

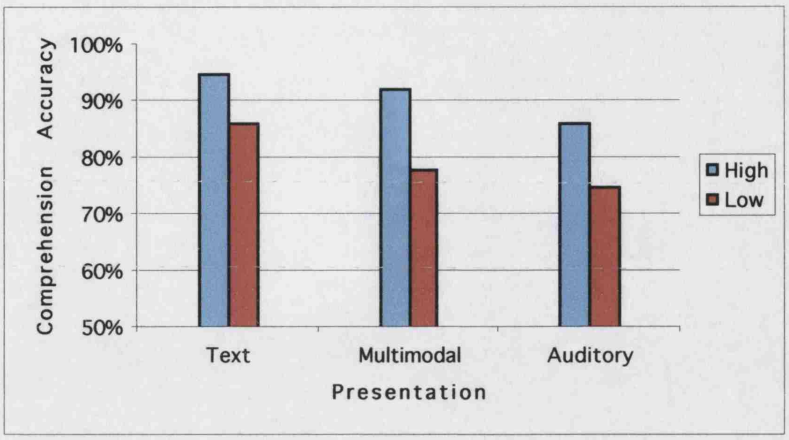


Table 8.4 and Figure 8.7 demonstrate that, consistent with the predictions for long-simple sentences, high span subjects showed higher comprehension rates than low span subjects ($F(1, 30) = 15.504$; $p < .01$); (c.f., P 2.2.2 and P 3.2.2). Also, the higher load imposed by these sentences produced differences in comprehension rates between the different presentation conditions ($F(2, 29) = 7.854$; $p < .01$). However, the source of this main effect is only compatible with the experimental prediction concerning speech-based systems. For this level of load, users were expected to benefit from visual text as an added modality in the multimodal condition (GL 3.2). The 4% advantage found for the multimodal condition over the auditory condition was powerful enough to support this assumption ($F(1, 30) = 5.050$; $p < .04$); (c.f., P 3.2.1). Furthermore, as predicted, the difference between these presentation conditions was evident for both low and high span subjects, as span and multimodality failed to interact for the auditory-multimodal contrast ($F(1, 30) = .036$); (c.f., P 3.2.3). Subjects' reports support the advantage found for the multimodal condition over the auditory condition. When asked which presentation best supported the longer test sentences, only 12.5% of the subjects suggested the auditory condition. On the other hand, 47% of the subjects suggested the multimodal condition.

In contrast to this multimodal facilitation, reading a long sentence whilst listening to it being spoken was hypothesised to selectively impair the comprehension rates of low span subjects relative to just reading the sentence (c.f., P 2.2.3 of GL 2.2). These subjects were assumed to have insufficient capacity to accommodate the storage and processing demands of long right-branching sentences and the coordination of processing between modalities. High span subjects were expected to successfully coordinate processing between the two modalities. Results show however no interaction between span and presentation for the visual-multimodal contrast ($F(1, 30) = .352$). Speech as an additional modality did not impair processing of the visual text for either span group and as a result, no

multimodal interference effect was evident (F (1, 30) = 1.835); (c.f., P 2.2.1)[83]. Moreover, the preference data, presented in Appendix F, suggests a perceived *advantage* of multimodal presentation over visual presentation for the longer test sentences. Only 25% of the subjects appear to suggest that the visual condition best supported these sentences, contrasting with the 47% expressed preference for the multimodal condition[84]. These findings contradict the multimodal interference in comprehending long-simple sentences that was identified for low span subjects in experiment 3 and hence, fail to validate GL 2.2. Rather, they imply that speech can be added to a static-durable visual display of a long-simple sentence without impairing user performance regardless of span.

On average, the analysis yielded a significant effect of verbal WM capacity; high span subjects performed significantly better than low span subjects (94% vs. 84%); (F (1, 30) = 20.929; $p < .01$). However, averaged across sentence length, the interaction between presentation and span failed to reach significance (F (2, 29) = .573); comprehension rates of all users were affected equally by the different modes of presentation. Furthermore, in spite of the fact that 2 clauses sentences yielded higher comprehension rates than 3 clauses sentences (92% vs. 85%), (F (1, 30) = 18.679; $p < .01$), the manipulation of sentence length was not strong enough to produce the expected interaction between length and span in this study (F (1, 30) = .000). Both sentence lengths imposed a higher load on low span subjects than on high span subjects. On the other hand, differences in sentence length were capable of producing the expected interaction between length and presentation (F (2, 29) = 4.628; $p < .03$): no effect of presentation for 2 clauses sentences and a significant effect of presentation for the longer 3 clauses sentences. Simple main effects yielded an effect of presentation in the auditory-multimodal contrast (F (1, 30) = 5.050; $p < .04$), no effect of presentation in the visual-multimodal contrast (F (1, 30) = 1.835) and an effect of presentation in the visual-auditory contrast (F (1, 30) = 12.848; $p < .01$). Table 8.5 and Figure 8.8 display the CR means created by these variables.

Table 8.5

Applied Study. Mean Comprehension Rates (CR) for all Presentation Conditions (%): By Sentence Length (Standard Errors in Parentheses)

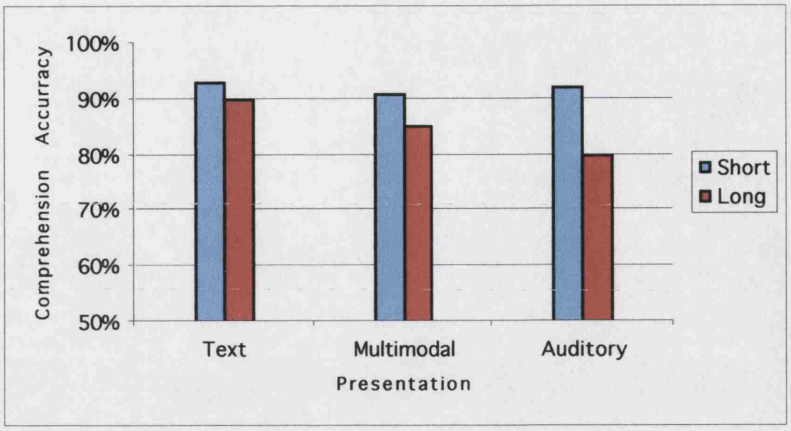
	Presentation By Length		
	Visual	Multimodal	Auditory
2 Clauses	93%	91%	92%
	(1%)	(2%)	(2%)
3 Clauses	90%	85%	80%
	(2%)	(3%)	(2%)
Mean	92%	88%	86%

[83] The added speech did not impair processing of the visual text in the visual-multimodal contrast in spite of the larger numerical difference relative to the (significant) auditory-multimodal contrast. The reason is some large deviations from the mean of individual results in the visual-multimodal contrast.

[84] Note that this difference was not analysed statistically.

Figure 8.8

Applied Study. Mean Comprehension Rates (CR) for all Presentation Conditions (%): By Sentence Length



Finally, the main effect of presentation reached significance as expected ($F(2, 29) = 5.854$; $p < .01$). The source of the effect is the 6% advantage of the visual condition (92%) over the auditory condition (86%) in which storage and processing demands were the highest ($p < .01$). The average differences between each of the unimodal conditions and the multimodal condition (88%) were too subtle to reach significance.

Response times

The time it took to comprehend each sentence was collected for each subject. These values were analysed in a repeated-measure analysis of variance. Again, span formed the between-subjects variable while length and mode of presentation formed the within-subjects variables[85]. Response times for 2 clauses sentences (2559 ms) were significantly faster than those obtained for 3 clauses sentences (2726 ms) ($F(1, 30) = 8.845$; $p < .01$) complementing the main effect of length found for the CR measure. Also, the main effect of presentation reached significance (F Huynh-Feldt (1.743, 52.305) = 4.088; $p < .04$). The source of the effect is the superiority of the durable over the transient presentation forms: response times in the auditory condition (2806 ms) were significantly slower than those of the visual condition (2537 ms); ($p < .03$) and marginally slower than those of the multimodal condition (2584 ms); ($p < .08$). These effects enhance the corresponding effects found for the CR

[85] An exploration of the response time data revealed that the sub-conditions created by the span, length and mode of presentation variables did not exhibit perfectly normal distributions. The assumption of normality was met for 7 out of 12 sub-conditions, although all distributions were positively skewed. Furthermore, the assumption of homogeneity of variance was met for all sub conditions. It was decided to conduct an analysis of variance over these values. Note however that the assumption of sphericity was not met for the *presentation* factor (Mauchly's W = .757, Approx. Chi-Square = 8.064; $p < .03$). An adjustment value, the Huynh-Feldt epsilon, was needed for multiplying the numerator and denominator degrees of freedom for the *presentation* factor in the *F* test.

measure but do not shed a new light on the findings described above. No other effect reached significance.

8.2.5 Discussion

Validation of GL 2.1 & 3.1

The results fail to validate GL 2.1 and GL 3.1 for short-simple sentences. Specifically, the significant effect of span found for the 2 clauses right-branching sentences challenges the revised MMUM: if the 2 clauses right-branching sentences were demanding enough to show differences between performance of high and low span subjects, why was an effect of presentation for low span subjects not seen? As noted earlier in this chapter, the revised model suggests that when sentence load increases, users rely on the durability of the visual text. At the very least, this reliance of low span subjects on the durability of the visual text should have yielded higher comprehension rates of 2 clause right-branching sentences in the visual relative to the auditory condition. Yet, the 4% difference, found between these conditions for low span subjects was not sufficient to produce the visual superiority effect (see Table 8.3 and Figure 8.6). With the absence of variations in the strength of context in the different presentation conditions, explanations for this finding may concern the features of the test sentences and of the presentation conditions, rather than aspects of the subjects or the task. Two complementary explanations are suggested to account for this finding (i) the inclusion of proper nouns and (ii) the use of intonation and prosody.

The inclusion of proper nouns: In the previous studies, noun phrases consisted of *common nouns*[86]. Determining which noun is the agent of which verb could be informed by subjects' mental model of the world; by pragmatic associations between nouns and verbs. However, because pragmatic associations between nouns and verbs are independent of syntactic relationship, this kind of information was sometimes misleading and could not replace syntactic processing. Evidence for the use of pragmatic information was provided by the *negative* relationship between pragmatic complexity in the previous studies and subjects' comprehension rates; the higher the pragmatic complexity value of the sentence, the lower was the CR. The effect was stronger for complex than for simple sentences, indicating a greater reliance on pragmatic information as complexity increased. In this study, *proper nouns*[87] replaced the use of common nouns to make the sentences more natural. The first two proper nouns in each test sentence were names of people, each of which could have been either the agent or the theme of the first predicate in each test sentence. As a result, pragmatic complexity values increased. However, as the selected names were arbitrary and free of pragmatic associations, the significance of pragmatic cues has diminished; results of an item analysis suggest that assigning verbs to names when processing the 2 clauses sentences was independent of subjects'

[86] Common nouns refer to something or someone as a member of the set of similar things. For example, a *cat* is a member of the set of all cats.

[87] Proper nouns usually refer to a particular, named person or thing (e.g., names of specific people, trade names, newspaper and magazine titles).

mental model of the world as no relationship between pragmatic complexity and comprehension rates was identified for these sentences (F (1, 20) = 2.133). It is suggested that the use of proper nouns with this fully predictable syntactic structure might have decreased the reliance on syntactic analysis and increased the use of strategies in all presentation conditions: assigning verbs to nouns might have become an automated task rather than part of the sentence comprehension process in this study. In support, subjects' reports suggest that the sentences were not processed for comprehension but rather for recall. Regardless of their span status, 78% of the subjects claimed to have focused on the order in which the names and the verbs were given. Of these subjects, 40% described focusing attention on the second name and its following verb (see Appendix F, Other strategy). This second name strategy enabled subjects to eliminate false noun-verb combinations with greater ease, with the result of equating performance of low span subjects in the auditory and the visual conditions.

The use of intonation and prosodic cues: This explanation concerns the addition of intonation and prosody to the speech output in this study. As noted in Chapter 2, prosody and intonation (or the rhythm and melody of the sentence) provide rich cues to the sentence syntactic structure. The natural speech used in this study may have facilitated syntactic parsing, by emphasising focal words and by marking clear syntactic boundaries. However, since the above explanation suggests that the test sentences were not processed for comprehension but rather for recall, it does not seem likely that it is the facilitation of syntactic parsing by prosody and intonation that accounts for the equal performance in the unimodal presentation conditions. A more probable explanation is that relying on intonation and prosodic cues complemented the recall strategies described above. This explanation proposes that intonation and prosodic cues facilitated the retention of word-order information by the phonological sub-system. Reference to these phonological representations assisted the recall of word-order information and hence, the assessment of noun-verb combinations specified by the comprehension statements.

In summary, the inclusion of proper nouns, intonation and prosody may have promoted recall strategies in this study. This was sufficient to compensate for the higher load imposed by the 2 clauses right-branching sentences on low span subjects, thereby cancelling processing differences in the unimodal conditions. The applied study aimed to test the validity of the guidelines under less sterile situations. It is possible however that relaxing previous laboratory constraints reduced the possibility of providing a coherent explanation for this level of load[88]. Furthermore, it appears that the 2 clauses right-branching sentences were not suitable for the validation of GL 2.1 and GL 3.1; their length of 11 to 16 words was too demanding to produce the pattern of results predicted by the MMUM for short-simple sentences. On the other hand, the findings for this level of load do not contradict the general suggestion that a multimodal presentation can substitute for a unimodal presentation of a short-simple sentence without increasing processing load. On the contrary, they

[88] Although this explanation refers to the lack of statistical difference between the unimodal conditions, it could also apply to the equality between the visual and the multimodal conditions. However, it is not assumed that low span subjects experienced multimodal interference at this level of load. Rather, it is suggested that the inclusion of proper nouns, intonation and prosody improved performance in the auditory condition, equating between this and the two other conditions.

suggest that the statement might even be valid for medium length sentences of a low syntactic complexity value. The use of recall strategies in this study calls however for further studies. These are needed in order to confirm the absence of multimodal interference for low span subjects under natural processing conditions of medium length sentences and in order to generalise such findings for shorter sentences of a similar syntactic complexity value.

Validation of GL 2.2 & 3.2

Results fail to show a selective multimodal interference in relation to the visual condition for low span subjects and hence, fail to validate GL 2.2. Furthermore, the preference data even shows a stronger preference for multimodal presentation (47%) over visual presentation (25%) for long-simple sentences. Taken together, these findings contradict the multimodal interference in comprehending long-simple sentences that was identified for low span subjects in experiment 3. How can this inconsistency be addressed? The 3 clauses test sentences used in this applied study share the same right-branching structure and the same number of clauses with the simple sentences of the previous studies. If processed for comprehension, they should have placed similar storage and processing demands on low span users and produced a similar speech interference effect. However, it appears that similar to the shorter test sentences, subjects tried to process them for recall. The use of proper nouns with a fully predictable syntactic structure has probably decreased the significance of pragmatic cues also for these sentences, as no relationship between pragmatic complexity and subjects' comprehension rates was found at this level of load ($F(1, 20) = 1.114$). Given that these 3 clauses right-branching sentences were not processed for comprehension, but that their key words were stored for later recall, then their processing demands were qualitatively different to those of the right-branching sentences used in previous studies. Under these circumstances, no speech interference should have been evident for low span subjects. Unfortunately, it is not certain that recall strategies were exclusive to this study since interviews were conducted in this study alone; similar strategies could have also taken place in the previous experiments. However, the use of pragmatic inferences in the previous experiments implies that subjects tried to comprehend the sentences; had they not tried to comprehend the sentences, pragmatic complexity would not have correlated negatively with the comprehension rates. Thus, the same extent of strategic behaviour couldn't have taken place when common nouns were used.

Finally, intonation and prosody might have complemented the recall strategies for this level of load. As suggested earlier for the 2 clauses right-branching sentences, the cues that were absent in the previous experiments might have been useful in the retention and recall of word-order information, thereby complementing the recall strategies described above. This might have made multimodal presentation more preferable to visual presentation for this level of load, in spite of the absence of a difference in the performance data. In summary, these two explanations suggest that the processing demands of the 3 clauses right-branching sentences were different in nature in this study than in the previous experiments. Low span subjects might have experienced some speech interference when trying to memorise the relationships between the nouns and the verbs but this was not strong enough to be realised by the data.

In contrast to the failure to validate GL 2.2, the results do validate GL 3.2 for long-simple sentences. Performance rates in both the visual and the multimodal conditions were higher than those in the auditory conditions for both groups of subjects, showing that for this level of load all users benefit from the presence of a durable visual text (see Table 8.4 and Figure 8.7). This can be taken to indicate that for speech-based systems, in which speech serves as the primary presentation channel, adding a static-durable visual text improves performance for long simple sentences. It seems that the demands in the auditory condition were those least affected by the use of recall strategies in this experiment, since a transient presentation makes it more difficult to extract order combinations when the sentence places a high processing load on the user. The sole reliance on intermediate computational products during the processing of long right-branching sentences in the auditory condition has probably increased the difference between this and the other presentation conditions

The preference data supports this multimodal superiority over the auditory condition, as only 12.5% of the subjects preferred the auditory condition in comparison to 47% who preferred the multimodal condition for the longer test sentences. It seems that regardless of their span, subjects benefited from the presence of the two channels in the static-durable multimodal condition. Specific processing strategies varied: 72% of the subjects reported consciously slowing down their reading pace to synchronise it with the speech rate (see Appendix F, Conscious synchronisation). However, even for these subjects, this conscious synchronisation was mixed with other strategies as 56% of all subjects reported that they used the visual channel to re-read parts of sentences in the multimodal condition (see Appendix F, 2nd reading). Some subjects scanned the sentences ahead and picked up the key words before they were spoken. Subsequently, when they heard the key words, they focused on them for the second time. Other subjects synchronised between the two channels until the middle of the sentence and then read ahead of the speech and regressed back to recollect the sentence's key words (see Appendix F, Multimodal strategy). This implies that neither the inclusion of proper nouns, nor the use of intonation and prosody could have made up for the higher load imposed by the 3 clauses sentences. For this level of load, the durability of the visual text was more important. The flexibility in processing the visual information in the static-durable multimodal condition might have supported more effective comprehension strategies, but may have also made it easier to extract word-order combinations, thereby facilitating the use of recall strategies in this study. The creative use of strategies by subjects in the multimodal condition is reflected in their clear preference for the multimodal condition in this study.

Conclusions

The results of this study failed to validate GL 2.1 and GL 3.1 for short-simple sentences. The selected 2 clauses right-branching sentences were too demanding to confirm the claim that a multimodal presentation can replace a unimodal presentation of short-simple sentences without increasing processing load. Furthermore, they did not enable an examination of the possibility that for low processing load levels, multimodality improves the overall satisfaction of all users relative to unimodal presentation of such sentences. In order to validate these claims, the study should be

repeated with a stronger manipulation of sentence length. Test sentences should consist of single clause sentences and 3 clauses right-branching sentences. Since comparison can involve only the first clause of such sentences, filler items should probe for the second and the third clauses. This will establish subjects' expectations that they will be asked about these clauses so that they need to attend them.

Results also failed to validate GL 2.2 for long-simple sentences in medium to large display devices. The inclusion of proper nouns and the added intonation and prosody were suggested to explain the increased use of recall strategies in this study, which made the multimodal interference effect that was identified for low span subjects in the previous studies disappear. As the selected names were arbitrary and free of pragmatic associations, the significance of the pragmatic cues was diminished and subjects processed the fully predictable syntactic structures for recall, focusing on the order of the proper nouns. Intonation and prosody might have supported this retention strategy: the cues that were absent in the previous experiments might have been useful in retaining word order information, thereby complementing the schematic mode of processing in this study.

To eliminate recall strategies, there is a need to add test sentences of different sentential structures, even if proper nouns or intonation and prosody are used. The reason is that in principle, neither the use of proper nouns nor the use of intonation and prosody promotes processing for recall under normal reading/listening conditions. The inclusion of various syntactic structures may encourage processing sentences for comprehension rather than for recall. If, following the addition of further syntactic structures, results replicate the multimodal interference effect for low span subjects, this would support the account that the absence of such an effect in this study stemmed primarily from the combination of a fully predictable simple structure with proper nouns. This combination should not be used to assess processing for comprehension, as responses will always involve recall strategies. Alternatively, if following the addition of further syntactic structures, results fail again to replicate the multimodal interference effect for low span subjects, then the role of intonation and prosody should be assessed separately. Intonation and prosody were suggested to have complemented the recall strategies used in the comprehension task of this study. However, their role in multimodal sentence processing might be more important than this. As noted earlier, intonation and prosody may facilitate syntactic parsing under natural processing conditions by emphasising focal words and by marking clear syntactic boundaries. As a result, low span users may have enough capacity to successfully accommodate the (now lower) storage and processing demands of long-simple sentences and the coordination of processing between modalities. Such an account could provide an important qualification for GL 2.2. It would imply that for long-simple sentences, natural speech with intonation and prosody could be added to a static-durable visual text while synthetic speech, with no intonation and prosody, should not. Without intonation and prosody, low span users cannot accommodate the storage and processing demands of long-simple sentences and the coordination of processing between modalities.

Finally, despite the above mentioned shortcomings, the applied study enabled validation of GL 3.2 for long-simple sentences presented by speech-based systems[89]. All measures converged into a clear advantage, when presenting long-simple sentences, of adding a static-durable presentation of text to systems which are primarily speech based:

- Comprehension rates of long-simple sentences were higher in the multimodal condition than in the auditory condition (the latter was the least affected by the use of strategies in this study).
- Response times in the auditory condition were marginally slower than those in the multimodal condition (the effect was not exclusive to the 3 clauses right-branching sentences).
- Users' reports indicated a clear preference of multimodal presentation over an auditory presentation for long-simple sentences.

Overall, this suggests that while the experimental design could be improved, the chosen approach was able to validate specific assumptions regarding the variation in the user's cognitive cost in response to different configurations and thereby to inform the design process of multimodal user interfaces.

[89] The advantage found of the multimodal and visual-only conditions relative to the auditory condition, taken together with absence of a difference between the visual-only and multimodal conditions, could be interpreted as simply an advantage of the presence of visual text over speech. This interpretation may imply that when users can read, they will tend to ignore any redundant speech output. Therefore a designer who was free to choose between the three forms of presentation, could not - on these results - be advised to choose other than the visual-only condition. However guideline 3.2 is specifically targeted at speech-based systems, for example, telephonic information systems, public address systems and the like, where the speech channel is given and the designer is considering whether there would be any advantage to the user in the addition of a redundant visual text.

Chapter 9

Conclusions

This chapter first summarises the research conducted in this thesis and then describes the revisions made to the assumptions of the MMUM as a result of its experimental work. A discussion regarding the contributions and limitations of the work follows, evaluating the research process and its products. Possible directions of future work are outlined last.

9.1 Summary of the dissertation

This research attempted to understand when and how speech can be combined with visual text to facilitate the user's processing and comprehension of sentences. Both theoretical and empirical methods were used to examine user's processing cost as a function of linguistic complexity and of memory demands placed by various multimodal presentation forms. The linguistic unit of this research was the sentence. The study did not attempt to cover contextual and discourse issues and was not concerned with embodied or situated tasks. Another scope decision was to address only fully redundant multimodality: the presentation of completely identical contents in both modalities. Understanding redundant multimodality would then form the basis for future studies of other forms of multimodality that were not assessed by this work (e.g., complementary-multimodality, complementary-redundant multimodality and partially-redundant multimodality).

At the theoretical front, the research made explicit both the central features of multimodal sentence presentation and the critical structures and processes involved in multimodal language processing. Two entities were presented:

- The MMDS - a theoretical space in which all types of redundant multimodal user interfaces can be found. The dimensions of the design space include (i) aspects of the verbal content (as expressed by linguistic complexity), (ii) aspects of presentation (as expressed by forms of media realisation and media coordination), and (iii) user cost. The model characterises sentential complexity (taking into account sentence length as well as the syntactic, semantic and pragmatic complexity properties of different sentences) and specifies multimodal presentation forms (a function of the *dynamism* and the *durability* of the visual text.)
- The MMUM - a cognitive model that characterises the structures and processes underlying multimodal language processing, including the supervisory attentional mechanisms that coordinate the processing of language in parallel modalities. Considering the verbal WM capacity of the user, the MMUM takes as input specifications of the content and its representational form and returns as output to the MMDS a statement of the expected user cost.

At the empirical front, the work enabled to validate through controlled studies with users, specific assumptions regarding the variation in the user's cognitive cost in response to different contents that are presented in different multimodal configurations. Multimodal configurations consisted of all combinations created by the dynamism and durability factors, as specified by the MMDS: dynamic-transient, dynamic-durable and static-durable multimodal presentations.

The first experiment, presented in Chapter 5, used the static-durable multimodal configuration to investigate the model's primary assumption of transition from a facilitatory synchronous processing, when the user is able to coordinate the processing of multiple modalities, to an interfering asynchronous processing, when this coordination fails. The transition is due to the accumulation of linguistic demands within a sentence. To examine this assumption, two modality-based conditions were used to present right-branching and doubly-embedded sentences with the same meaning: visual-based and auditory-based. In the visual-based conditions, a static-durable visual text was compared with a (static-durable) multimodal visual-attend presentation (MMVA) whereas in the auditory-based conditions, spoken sentences were compared with a (static-durable) multimodal auditory-attend presentation (MMAA). To manipulate attention, the experiment used a word-category monitoring-task with target words located in one modality. By placing the target words in either early or late positions in the test sentences, the experiment sought to demonstrate how user's cognitive cost changes through the course of processing a sentence. As well as monitoring for the target words, subjects had to perform a comprehension task for each sentence. Contrary to the predictions of the MMUM, the word-monitoring times failed to demonstrate a transition from a facilitatory synchronous processing to an interfering asynchronous processing under accumulating load conditions. Monitoring times for late target words were unaffected by sentence complexity and showed no evidence of synchronisation between the auditory and the visual modalities in this static-durable multimodal presentation. Furthermore, the comprehension measure indicated that whereas for simple sentences the addition of static-durable text improved performance relative to the speech condition, for syntactically complex sentences adding both a static-durable text to the speech channel and adding speech to the static-durable visual text channel reduced user cost. Two alternative explanations were provided to account for this result. A substantive explanation, the durability account, suggested that the availability of the durable visual presentation for a further recollection of information enabled subjects to visually regress to earlier sentential components while simultaneously attending to the spoken continuation of the sentence. This enabled them to advantageously switch attention between modalities so as to perform a delayed assignment of thematic roles across them. The alternative methodological explanation proposed that the observed facilitation was an artifact caused by qualitative variations in the monitoring task. This account suggested that with the removal of the word-monitoring task, the added speech would impair processing of complex sentences, as originally postulated by the MMUM.

The second experiment, presented in Chapter 6, aimed to decide between these two alternatives by manipulating the durability of the visual text. Two dynamic presentation techniques were used: a dynamic-transient format and a dynamic-durable format. Multimodal presentation consisted of presenting the two visual text conditions with redundant coupled speech. Subjects were required to

comprehend the same sentences that were used in experiment 1, this time without having to monitor for target words. In terms of the MMDS, the use of the dynamic-transient multimodal presentation enabled a study of the effect of multimodality in itself on user cost given variation in syntactic complexity, whereas the dynamic-durable multimodal presentation enabled an examination of the added effect of the durability of dynamic visual-text on user cost. Results were in line with the predictions made by the MMUM: the added speech was found to impair the comprehension of complex sentences and (when calculated across the verbal WM capacity groups) the comprehension of simple sentences was unaffected. Specifically, the magnitude of interference was strongest for complex sentences presented in the dynamic-durable form. It was suggested that under a dynamic-durable multimodal presentation of complex sentences, the incompatibility of the spoken information that occurs with visual regressions (following the breakdown of normal processing) produces an interference effect at all levels of processing. In addition, individual differences in verbal WM capacity were found to mediate user cost in this experiment: high capacity subjects were more resistant to multimodal interference than low capacity subjects regardless of the complexity of the sentences. Furthermore, for low span subjects, a significant speech interference effect was found in comprehending long-simple sentences presented in a dynamic-transient form (see combined analysis of experiments 2a and 2b). Although more subjects were needed to demonstrate a similar effect for the dynamic-durable presentation form, this provided some indication that the incompatibility of spoken information with recollected computational products might take place even when the assignment of thematic roles is immediate. Subsequently, multimodal interference seems to depend on the level of processing load experienced by the user.

An extension of the durability account was suggested to address the inconsistency in comprehension of complex sentences between experiments 1 and 2: the facilitation found for the static-durable multimodal format in experiment 1 vs. the interference found for the dynamic-durable multimodal format in experiment 2. This explanation proposed that multimodal facilitation in comprehending complex sentences might be limited to the static-durable presentation format in which processing of the visual information is user-paced and does not take place when it is machine-paced. The static-durable multimodal presentation allowed the users to scan and to skim the visual text, including making regressive eye-movements to previously processed portions of text. In contrast, the dynamic-durable multimodal presentation made regressive eye-movements difficult to carry out and as a result, the ability of the subjects to switch attention between modalities and to perform a delayed assignment of thematic roles across them was impaired. The third experiment, described in Chapter 7, aimed to examine the validity of this account using the static-durable multimodal presentation and the dynamic-durable multimodal presentation. Comparing these two multimodal techniques with their solely visual counterparts enabled the determination of the role of visual-processing control in processing multimodal information and by so doing, to decide between the extended durability account and the methodological account of the results of experiment 1. The use of these two durable presentation techniques also allowed an investigation of the effect on user cost of *coupling* redundant visual and auditory words, given variations in syntactic complexity. The MMUM assumed that presenting visual and auditory words together at the same rate facilitates the recoverability of synchronous processing under high processing demands. This recoverability of the dynamic-durable

multimodal format was predicted to alleviate processing load in comparison to the static-durable multimodal format that lacks this property. Consistent with the assumptions made by the MMUM, multimodality was found to impair sentence comprehension under high processing demands. This multimodal interference was not affected by the dynamism of the visual text, implying that the multimodal facilitation in comprehending complex sentences that was found in experiment 1 resulted from methodological limitations in the design of that dual-task experiment. The extended durability account was therefore rejected. Moreover, high visual-processing control was found to be more important than the recoverability of synchronous processing in alleviating processing cost under increasing load conditions, suggesting that the importance of coupled (synchronous) presentation in recovering from asynchronous processing was overestimated by the MMUM. Finally, high capacity subjects showed stronger resistance to speech interference in the simple condition than low capacity subjects. In the complex condition, the capacity variable failed to distinguish between comprehension rates in the unimodal and the multimodal conditions and between magnitudes of interference in the dynamic-durable and the static-durable multimodal modes of presentation (possibly due to a floor effect in the dynamic-durable multimodal condition).

In summary, the results of experiments 2 and 3 imply that long-simple sentences should be presented without additional speech, if one aims to accommodate all users. These findings were translated to a set of guidelines for effective multimodal presentation of sentences in Chapter 8, all of which comply with the limitations of low span users. The first set of guidelines referred to adding speech to visual text in systems in which the visual text is the primary channel and the second set referred to adding visual text in speech-based systems. All guidelines took into account the linguistic complexity of the presented sentences and the size of the visual display.

A final study, also reported in Chapter 8, sampled some of these guidelines with the aim of examining their validity in an applied setting. The setting simulated a hand held, palm sized email system that displayed static-durable messages on a screen, read them out to the users or combined the two modes of presentation by means of a non-coupled static-durable multimodal presentation. A realistic context was assumed for the email messages. Aiming to make the messages as natural as possible, only simple sentences were used; all email messages were a mix of short-simple priming sentences followed by either short (2 clauses) or long (3 clauses) right-branching test sentences. It was found that the selected 2 clauses right-branching sentences showed differences between performance of high and low span subjects and were therefore unsuitable for the validation of guidelines 2.1 and 3.1 provided for short-simple sentences. In addition, results failed to replicate the multimodal interference found for low span subjects comprehending long-simple sentences in both experiments 2b and 3. Contrary to guideline 2.2 that was provided for long-simple sentences in medium to large display devices, results implied that speech could be added to a static-durable visual display of a long-simple sentence without impairing user performance. Rejecting guideline 2.2 might be premature however, since the inclusion of proper nouns, intonation and prosody was suggested to promote recall strategies in this study, so the absence of span-based differences in comprehending long-simple sentences might have been artificial. Finally, results indicated a clear advantage of the static-durable multimodal presentation over the dynamic-transient auditory presentation for long-

simple sentences, thereby validating guideline 3.2 provided for such sentences in speech-based systems. While these results highlighted the need for experimental control, the selected approach was able to validate specific assumptions regarding the variation in the user's cognitive cost in response to different configurations and thereby to inform the design process of multimodal user interfaces.

9.2 Validation and refinement of the model

Throughout the experimental work, a consistent effort was made to validate and refine the model's assumptions. This section summarises the revised assumptions of the model, following the set of studies outlined above.

The overall pattern of results obtained in this work suggests that the assumption that the SAS supervises coordination by synchronisation between the visual and the auditory channels so as to maximise the multimodal activation of the cross-modal sub-systems was over simplified. The only attempt to directly examine the presence of synchronous processing took place in experiment 1 and failed to provide evidence for its existence. It was suggested that the priority given to the word-monitoring task over the comprehension task in this experiment made multimodal sentence processing anomalous and that under natural processing conditions, some form of synchronous processing might take place. Experiment 2 approached the investigation of synchronous processing from a different angle, as the visual and the auditory words were presented together at the same rate in both multimodal conditions. Results showed that under high load conditions, presenting the visual and auditory words together at the same rate did not assist processing. The experiment identified a significant speech interference effect for low span subjects comprehending long-simple sentences in the dynamic-transient conditions (see combined analysis of experiments 2a and 2b), but no effect in the dynamic-durable conditions (presumably due to low experimental power in experiment 2b; the same configuration yielded a significant effect for low span subjects in experiment 3). For the complex sentences, the coupled presentation formats severely impaired performance for all subjects (see experiment 2b). A more sophisticated form of coordination between modalities was proposed: when the visual and the auditory words are presented together at the same rate, multimodal activation of the cross-modal sub-systems reduces processing cost as long as processing load remains low for a particular user. When processing load increases, either due to syntactic complexity or (for low span subjects) due to sentence length, processing may breakdown and so users may attempt to retrieve previously processed information. As long as the available resources are sufficient, processing may proceed with no interference. However, when resources are low, the re-activation of intermediate computational products (in a dynamic-transient presentation) or the recollection of visual information (in a dynamic-durable presentation) will be impaired by concurrent speech.

Two other related assumptions that were refuted by the experimental work were (i) that the control over visual processing does not assist performance when processing complex sentences and (ii) that the recoverability property of synchronous processing is of a greater importance than the property of visual-processing control when processing demands are high. In brief, the MMUM claimed that following the production of regressive eye-movements, synchronous processing is easier to restore in

the dynamic-durable multimodal format than in the static-durable multimodal format through refocusing attention on the "leading edge" of the visual display. This affordance to restore synchronous processing was expected to produce weaker speech interference under dynamic-durable multimodal presentation than under static-durable multimodal presentation of demanding sentences. Experiment 3 put these assumptions to the test, comparing the comprehension of right-branching and doubly-embedded sentences in these two multimodal formats. In contrast to the predictions, the flexibility in processing the visual information in the static-durable presentation of doubly-embedded sentences facilitated the repair process of the syntactic processing breakdown, enabling a better assignment of thematic roles between sentential constituents. Moreover, such a high visual-processing control was found to be more important than the recoverability of synchronous processing in alleviating processing cost under increasing load conditions. The multimodal interference effect, found in comprehending doubly-embedded sentences (and right-branching sentences for low span subjects), was not affected by the dynamism of the visual text; presenting the visual and the auditory words together at the same rate did not reduce the interference effect in comparison to the static-durable conditions.

However, the most interesting assumption of the MMUM, that the SAS relies upon the same limited pool of resources used for sentence processing for its multimodal supervision functions, gained experimental support. With the exception of experiment 1, in which the dual-task conditions confounded the experimental results, multimodality was found to impair sentence comprehension under high processing demands in both experiments 2 and 3. It appears that when processing load increases, either due to syntactic complexity or (for low span subjects) due to sentence length, fewer resources are available to the SAS. Consequently, its ability to supervise coordination of processing between modalities is impaired and multimodal interference occurs. Thus, although multimodal sentence processing does not appear to consist of synchronisation between modalities, conflicting representations do raise a multimodal interference when resources are low. As noted earlier, this interference might take place at a semantic level of processing, as when background speech impairs performance of a concurrent reading task involving *different* materials (Martin et al., 1988) although an interference on a phonological basis cannot be ruled out.

Finally, the assumed relationships between verbal WM capacity, syntactic complexity and multimodality are yet to be proven. Individual differences in verbal WM capacity were expected to affect performance when the combination of the multimodal presentation technique and the linguistic complexity of the presented materials imposed a high processing load on the users. The higher the processing load, the more apparent the difference was assumed to be. The sentence length used in experiments 2 and 3 was capable of distinguishing between the comprehension rates of users with low and high verbal WM capacity for simple sentences; the added speech impaired performance of low span subjects and left high span subjects unaffected. However, in both experiments the span variable failed to statistically distinguish between the varying magnitudes of multimodal interference for complex sentences. The limited number of span subjects in each complexity condition and the poor screening method of span users might suggest that abandoning the assumed relationships between linguistic complexity and verbal WM capacity in the multimodal domain is premature.

However, it is also possible that the assumed relationship cannot be validated using excessively complex sentences for which all users experience non-normative processing. An intermediate level of complexity (e.g., 4Xint PLUs) might be needed to confirm these relationships.

9.3 Contribution of this dissertation

This dissertation adds knowledge to the field of cognitive engineering through corroborating modelling and empirical approaches as valuable methods for studying, understanding and improving the design process of multimodal user interfaces. Additionally, it makes three specific contributions: the development of the MMDS, the development of the MMUM and the formulation of a list of guidelines for effective multimodal presentation of sentences that vary in their linguistic complexity. These will be discussed next.

9.3.1 The development of the multimodal design space (MMDS)

The MMDS provides a modest contribution of research, as the dimensions it provides derive from different sources that were never brought together in previous research. The model uses Maybury's (1993) concepts of content selection, media allocation, realisation and coordination and structures presentation in the form of two dimensions, one for content and the other for form. User's processing cost depends upon the relationships between these dimensions. Although the distinction between content and form was made by Maybury's definitions, the effect that complexity of content has on the preferred presentation format was never discussed.

The characterisation of content proposed by the MMDS distinguishes between variations in linguistic complexity using objective means:

- Gibson's (1991) complexity metric was selected to determine the syntactic complexity of sentences.
- Sentence length, a widely used factor in readability formulas, was taken to qualify memory load.
- A simple pragmatic complexity metric was devised to determine the ease of interpretation based solely on a pragmatic reading of a sentence.

The characterisation of presentation combines Bernsen's (1994, 2001) concept of dynamism with the concept of text durability to define how the realisation of visual text determines possible modes of coordination between modalities in redundant multimodal presentation. Contrary to previous multimodal taxonomies, this characterisation makes processing cost of various verbal contents explicit for different multimodal formats. Processing cost, as predicted by the MMUM, was suggested to depend on (i) the memory demands imposed by the dynamism and the durability of the multimodal presentation type and (ii) the management of synchronisation between visual and auditory words in coupled and non-coupled redundant presentation forms. Making these factors explicit enabled this research to put them into test and to conclude the greater significance of the first factor in predicting the processing cost of various verbal contents. As suggested earlier, multimodal

sentence processing does not appear to consist of a simple synchronisation between modalities. More sophisticated forms of coordination were suggested to explain user's performance in this work.

In its current form, the MMDS is a space in which only redundant multimodal interfaces can be found; it does not represent other forms of multimodality such as complementary-multimodality, complementary-redundant multimodality and partially-redundant multimodality (see Chapter 1). In spite of its limited focus, the characterisations of content and presentation were successful in guiding a coherent set of usability studies, each with a different focus of research. These characterisations may also be useful in the investigation of other forms of multimodality, as the memory demands imposed by content and form are not specific to redundant multimodality. Specific examples will be discussed in section 9.5.

9.3.2 The development of the multimodal user model (MMUM)

The MMUM characterises how a person reads and listens to identical materials at the same time and how a single understanding of the input is formed. The development of this model provides a more substantial contribution of research, using Norman and Shallice's (1986) model of executive control to extend Just and Carpenter's (1992) capacity theory.

In the introduction of this thesis, three general approaches of user cognition were considered in terms of their suitability to provide a base upon which to build a cognitive model of multimodal language processing; all three approaches were found to be inadequate. As an alternative, it was decided to use Just and Carpenter's (1992) capacity theory. The use of this language processing framework, with a built-in mechanism for explaining capacity limitations, was considered optimal to guide the design of effective multimodal systems despite the absence of a mechanism to explain the separate processes of reading and listening and their convergence. The MMUM extends this work; the model provides a systematic account of individual and common mechanisms for speech and text processing, bringing into a single conceptual structure established theories of reading, of listening and of a-modal language comprehension. In this framework, Norman and Shallice's (1986) model of executive control is used to explain the control of language processing as well as the control of coordination between modalities. The model supports both automatic and conscious levels of multimodal processing control, the involvement of which is a function of both content complexity and its representational form.

The MMUM is still relatively basic. Section 9.2 described a number of changes to the model's assumptions, required by the experimental results. In addition, section 9.4.1 suggests how the development of more sophisticated linguistic complexity metrics could contribute to the elaboration of the model (e.g., in providing more complex predictions of user cost). In spite of its basic form, the formation of the model was necessary since, as noted in Chapter 1, previous explanations of multimodal language processing have been limited to single word presentation of speech and visual text. The MMUM attempted to explicate the processing of whole sentences. Changes of assumptions were to be expected when making such a large shift in the complexity of both content and form. In

addition, the fact that the model can change and be improved with advances in science is a positive thing. Overall, the MMUM proved useful in its utilisation of capacity limitations in multimodal language processing, providing refutable hypotheses of variation in the user's cognitive cost for different contents that are presented in different multimodal configurations.

9.3.3 The formulation and validation of guidelines for effective multimodal presentation of information varying in linguistic complexity

As suggested in Chapter 1, previous guidelines for the development of multimodal presentation provide no advice on how to present sentences that vary in their linguistic complexity with respect to cognitive constraints. These guidelines either refer to a single modality (e.g., Smith & Mosier's (1986) guidelines for speech output), or where referring to multimodal presentation (e.g., W3C organisation, Faraday & Sutcliffe (1997)), they may be concerned about form but not about content; no direct reference is provided for the linguistic complexity of the presented sentences. The guidelines provided by this work aim to determine optimal presentation forms of sentences that vary in their linguistic complexity. These guidelines account for the linguistic complexity of the presented sentences, for the memory demands of the primary presentation channel and for the verbal WM capacity of the user. Their ability to account for the processing limitations of low span users makes them particularly important for multimodal applications intended for the elderly since, as noted earlier, Just and Carpenter (1992) suggest that reduction in WM capacity occurs with ageing. The guidelines are also of relevance for the design of multimodal applications intended for the general population, since all types of users should be accounted for. Finally, the guidelines are equally applicable for conceptual and for procedural tasks. They are formulated in simple terms, making them easy to use by multimodal interface designers, and they stand alone, as no reference to the MMUM or to the experimental work is necessary for their understanding.

Due to their limited scope, the guidelines are incomplete in their ability to account for all types of multimodal user interfaces. First, they only account for fully redundant multimodal interfaces and second, they were derived using a single sentence length and only two linguistic structures that were chosen to represent different sentence complexity values (e.g., X_{INT} PLUs and $5X_{INT}$ PLUs). However, completeness was never a goal of this thesis, rather, it attempted to demonstrate the type of guidelines required to support the processing capabilities of users and the research that needs to be conducted in order to derive them. Furthermore, it has established a paradigm for deriving guidelines for multimodal interface design that can be used when researching other contents and other presentation forms.

Are the guidelines valid and true? The guidelines presented by this work may suffer from low external validity. The guidelines constructed for long-simple and long-complex sentences might be based on inaccurate interpretations of the experimental results, affected by ceiling and floor effects of the data. For example, the guidelines for visual-based systems suggest that high span subjects are resistant to speech interference when processing long-simple sentences. However, it is possible that for these subjects, concurrent speech actually *improves* performance for long-simple sentences. If,

having been exposed to a single long-simple structure in both experiments 2b and 3, high span subjects hit the ceiling in their comprehension rates in the unimodal conditions, this might have undermined the attempt to reveal a slightly higher comprehension rates in the multimodal conditions for these subjects. This suggests that multimodal facilitation might be found for high span subjects in a real-task situation involving various long-simple structures. Another example for a possible interpretation constraint is provided by the guidelines constructed for long-complex sentences in visual-based systems. These guidelines do not distinguish between differential levels of speech interference for different groups of subjects in different multimodal configurations. They simply suggest that for long-complex sentences, all users experience speech interference effect. However, as described earlier, the absence of a significant interaction between dynamism, multimodality and span for complex sentences in experiment 3 might have been affected by a floor effect. If high span subjects hit the floor in their comprehension of complex sentences in the dynamic-durable multimodal condition (as indicated by chance comprehension rates in this condition for all subjects), this might have defeated the attempt to reveal a greater speech interference effect for low span subjects in this condition[90]. The use of a less complex structure with which the parser can cope successfully (e.g., 4Xint PLUs) could have possibly yielded this pattern of results. In spite of these interpretation difficulties, the valid interpretations that: (i) adding speech to text impairs the comprehension of long-simple sentences for low span users, and (ii) adding speech to text impairs the comprehension of long-complex sentences for all users, are assumed to be critical for any design decision concerning multimodal presentation of such sentences.

The guidelines constructed for short-simple sentences (c.f., guidelines 2.1 and 3.1) may also suffer from low external validity, but for a different reason. These guidelines are based on minimal empirical evidence as none of the investigative studies (experiments 1 to 3) actually manipulated sentence length. The applied study attempted to test and improve the validity of these guidelines and those of long simple sentences (c.f., guidelines 2.2 and 3.2). This attempt had limited success. The study tested the validity of these guidelines in a representative setting, with added task context, intonation and prosody. Nevertheless, the manipulation used was not sensitive enough to validate guidelines 2.1 and 3.1 for short-simple sentences. The lack of sensitivity was not a result of the added task context. Rather, the 2 clauses right-branching sentences were too demanding as they produced differences between performance of high and low span subjects. They could therefore, not be used to confirm the claim that a multimodal presentation can substitute a unimodal presentation of short-simple sentences without increasing processing load. On the other hand, the claim has not been shown to be untrue. On the contrary, the findings suggested that the statement might also be valid for medium length sentences of a low syntactic complexity value. The use of recall strategies in this study calls however for further studies. These are needed in order to replicate the absence of multimodal interference for low span subjects under natural processing conditions of medium length

[90] In Experiment 2b, the absence of a significant interaction between durability, multimodality and span is not mediated by a possible floor effect. In this experiment, comprehension rates of complex sentences for high span subjects were higher than chance in all presentation conditions.

sentences and in order to generalise such findings for shorter sentences of a similar syntactic complexity value.

In addition, the absence of span-based differences in comprehending long-simple sentences failed to validate guideline 2.2 that was provided for long-simple sentences in medium to large display devices. In contrast to this guideline, results implied that speech could be added to a static-durable visual display of a long-simple sentence without impairing user performance. Again, the inclusion of proper nouns, intonation and prosody was suggested to promote recall strategies in this study so results might have been artificial and rejecting guideline 2.2 might be premature. Finally, results indicated a clear advantage of the static-durable multimodal presentation over the dynamic-transient auditory presentation for long-simple sentences, thereby validating guideline 3.2 provided for such sentences in speech-based systems. Overall, more research is needed to validate the suggested guidelines, with a strong emphasis on enhanced experimental control.

9.4 Limitations of this dissertation

This dissertation consists of three major parts. The first is concerned with making the features of both multimodal sentence presentation and multimodal language processing explicit. The second part covers the empirical work, examining through controlled studies with users specific assumptions regarding the variation in the user's cognitive cost in response to different contents that are presented in different multimodal configurations. Finally, in the third part, the experimental findings are translated to a set of guidelines that are sampled for a further validation in an applied setting. The three parts make up a complete story – starting with background, focusing on empirical studies and ending with prescriptive guidelines for effective multimodal presentation. While each part of the story could be refined and discussed further, a number of specific theoretical and empirical limitations of the study deserve special attention.

9.4.1 Theoretical aspects

Dissociating cognition from context

The difficulties of generating a theory of cognition that will support an effective bridging to artifact design are numerous. Some suggest that producing a useful theory of human cognition for HCI is an impossible goal, due to the complex and unpredictable nature of human behaviour (e.g., Landauer, 1991). For them, no general theory can provide a complete specification of user requirements, which could replace empirical evidence of a crude prototype tested with a handful of users or a small panel of experts. A more flexible attitude (e.g., Green, 1991) accepts that cognitive theory can develop hypotheses and models of the fine details of mental architecture and of specific architectural components, but in rarefied conditions that cannot be generalised into artifact design.

In the attempt to provide coherent explanations of multimodal language processing, this work used reductive methods, maximising experimental control in constrained laboratory settings. It could be claimed however that in its dissociation of cognition from context, the work sacrificed ecological validity. The investigative studies used unrealistic materials and did not address discourse issues (e.g., the effect of placing sentences within paragraphs). Moreover, they ignored the potential effects of domain knowledge, task difficulty and dialogue usability on processing sentence materials. Social context, emphasised by situated cognition theories (e.g., Lave, 1988), was not taken into account either. In defence, it is suggested that this research does not deny the significance of such high level factors when considering the processing of verbal information: multimodal sentence processing is not independent of contextual factors and these may overshadow fundamental cognitive processes. However, there is a real need to uncover fundamental aspects of multimodal sentence processing, a task that cannot be performed in an applied context, especially when no prior knowledge exists to inform the investigation of these aspects. Other things being equal, these fundamental aspects determine the ease with which messages presented to both modalities are understood. Questions of how isolated sentences are interpreted and used within context and of how sentence processing is affected by domain knowledge, task difficulty, usability factors and social factors are very important but secondary to the investigation of the fundamental aspects of multimodal sentence processing.

Yet, this work introduced some contextual factors in the applied study: a simple multimodal application was simulated, a realistic task context was assumed and priming sentences preceded the test sentences. On the other hand, even in this study, the task and the dialogue were designed to have a minimal effect on processing the test sentences and therefore, to enable the assessment of further aspects of multimodal sentence processing (e.g., sentence length). Time and experiments were limited, so the study had to compromise the wish to uncover fundamental aspects of multimodal sentence processing with the wish to validate guidelines derived from the controlled investigative studies.

In conclusion, examining activities within the larger context in which they occur is an important factor in designing user interfaces. However, this approach must be applied onto a mature domain of research, or one that, at the very least, possesses a good understanding of the phenomena in question. At this early stage of research into multimodality, the fundamental aspects of multimodal sentence processing could not be assessed in real-work situations, since these introduce various confounding explanations to the complex phenomena in question. Contextual factors could be assessed in future work, possibly by means of further controlled studies, adding factors one by one, to highlight each specific effect on multimodal sentence processing. As long as contextual factors are added to the experimental design in a controlled way, their inclusion can inform the investigation of the fundamental aspects of multimodal sentence processing and produce interesting and relevant data.

Using a rudimentary complexity metric

This research used Gibson's (1991) complexity metric to define syntactically simple and complex sentences. As the work progressed, the field of linguistics has itself grown and developed. The

primary representation of complexity was replaced by Gibson in 1998 with a more sophisticated measure with greater explanatory power. Contrary to the old metric, the new metric can explain why a doubly-embedded sentence is easier to process when an indexical pronoun is in the subject position of the most embedded clause, as compared with similar structures in which a proper name, a full NP or a pronoun with no referent is in that position. The new metric assumes that the referent of an indexical pronoun is already embedded in the current discourse. Since it is not considered to be a new discourse referent, it does not increase processing load. As noted in Chapter 2, this new metric has two components: a memory cost component and an integration cost component associated with keeping track of obligatory syntactic requirements. The memory cost component is quantified in terms of the number of syntactic categories that are necessary to complete the current input string as a grammatical structure. Moreover, this component defines what quantity of computational resources is required to store part of an input sentence. The second component, the integration cost, specifies the quantity of computational resources required to integrate new discourse referents into the structures built so far. In a remarkable relevance to this dissertation, these two components correspond to the storage and computation components of Just and Carpenter's (1992) capacity theory. Following the capacity theory, Gibson (1998) assumes that both linguistic integration processes and storage access the same limited pool of WM resources. This dependency affects integration time, as indexed by reading time: integration will take longer to achieve if some of the computational resources of that pool are used for storage as well as if the pool is smaller (i.e., for individuals with lower verbal WM capacity). Thus, similar to the old metric, the new account enables an assignment of complexity value to any sentence and predicts the exact word where processing breakdown will occur in complex sentences (see Chapter 3). Its additional aspects include providing a dynamic quantification of the level of activation throughout sentence processing for different span users.

Since the newer metric relates syntactic complexity to the language-processing model upon which the MMUM was built, it could be claimed that the newer metric should have been used to replace the old metric. However, in spite of the fact that the newer theory has a greater explanatory power than the 1991 metric and although it directly relates to verbal WM capacity, it was decided that the old metric is sufficient for the purpose of this work. The main reasons were both theoretical and practical:

- Both metrics provide similar characterisations of syntactic complexity for the sentences used in this study: the doubly embedded structure that was used extensively in this research always used a full NP for the subject position of the most embedded clause, and thus correctly reflects the unacceptability of $5X_{INT}$ PLUs structures.
- Both metrics do not account for the sentence length factor in determining linguistic complexity. According to the new metric, both the memory cost component and the integration cost component are influenced by the notion of locality rather than of length, so the criticism of the 1991 metric also applies for the 1998 metric. The number of nouns and verbs that a sentence contains is meaningless, unless these are located at open syntactic prediction points: when storing a partial input sentence, the longer a predicted syntactic category is maintained in memory the greater the cost. Also, when integrating new words into the structure, the greater the distance between an incoming word and the head or dependent to which it attaches, the greater the

integration cost. These costs relate only indirectly to sentence length, as syntactic categories will be maintained longer in memory only in long sentences.

- The assessment method of user cost in this work was incompatible with the index of user cost specified by the 1998 metric. Variations in integration times due to differences in verbal WM capacity required measurement of word processing times rather than of sentence comprehension that was used in this research.

In conclusion, the rudimentary 1991 metric used in this research can be replaced in future work by the more sophisticated 1998 metric. The newer metric may contribute to the elaboration of the MMUM in providing detailed predictions of integration cost for different users when processing different sentences in various presentation types. Future work could assess these detailed predictions using eye-tracking methods, to determine a moment-by-moment profile of multimodal sentence processing.

9.4.2 Empirical aspects

The use of repeated-measure design

In the investigative studies, each subject had to process sentences of one syntactic style in four presentation conditions. Presentation was varied across uniform blocks following the assumption that mixing the different presentation modes would disrupt performance. Consequently, the design enabled subjects the opportunity to acquaint themselves with the tasks, procedure, presentation conditions and the linguistic structure used, raising the prospects of practice effects (subjects get better at the task as time passes) and fatigue effects (subjects get tired as time passes). One way to minimise such effects was the employment of counterbalancing for the orderings of the presentation conditions. The aim was to spread any order effects evenly across conditions so that order effects would not be confounded with the experimental treatments. This involved the creation of four experimental versions for each complexity group in each experiment, each having a different order of the presentation blocks. Additional analyses were conducted to assess the success of this prevention method for each study. These included the order in which subjects performed the four presentation conditions as an additional between-subjects variable (called version number) to make sure that the experimental results were unaffected by order effects. Whereas this was found true for most of the results, the revised analysis of experiment 2b yielded some practice effects relating to the verbal WM capacity variable; an asymmetrical transfer (the effect upon B being preceded by A is different from the effect upon A of being preceded by B) was identified for low span subjects with respect to both the durability and the multimodality variables. Specifically, for the durability factor, when the durable conditions preceded the transient conditions, comprehension in the durable conditions was lower than comprehension in the transient conditions. For the reverse order, comprehension in the durable conditions was either lower or *higher* than the comprehension in the transient conditions. All effects only approached significance. Similarly, for the multimodality factor, when the multimodal conditions preceded unimodal conditions, comprehension in the multimodal conditions was lower than comprehension in the unimodal conditions. In contrast, when unimodal conditions preceded

multimodal conditions, the effect of multimodality did not reach significance. Although this interaction between span and multimodality only approached significance, this result indicates that practice might have contributed to the overall speech interference effect found for low span subjects in this experiment. One way to ameliorate the problem could have been to make sure that subjects were well practiced before starting the experimental session. Similarly, inserting effective rest periods between presentation conditions would have improved any sensation of fatigue, although no direct evidence for this effect was obtained in any of the revised analyses.

The use of excessively complex sentences

Two syntactic structures were selected to assess the effect of syntactic complexity on the user's ability to coordinate between visual and auditory information: the right-branching and the doubly-embedded. As noted in Chapter 5, these syntactic structures are widely used in psycholinguistic studies, due to their qualities of imposing differential demands on the parser while preserving the semantic relationships between the nouns and the verbs. The doubly-embedded structure was chosen to demonstrate the transition from a facilitatory synchronous processing to an interfering asynchronous processing in experiment 1. The monitoring task intended to show that the accumulation of linguistic demands within this complex structure imposes an excessive processing load on the syntactic and semantic systems and, as a result, slows down the computation of late-position target words by the lexical-access system. However, as described earlier, monitoring times for late target words were unaffected by sentence complexity. Even at this preliminary stage of research, it was suggested that the doubly-embedded sentences might have been too complex to enable the SAS to scale the allocation of attention between the lexical-access system and the syntactic and semantic systems in the predicted manner. However, the surprising finding that multimodal presentation facilitates the comprehension of these excessively complex sentences necessitated a further investigation of the effect using the same syntactic structure. Results of experiments 2 and 3 did not repeat this facilitation, as multimodality was found to impair processing under high load conditions. However, this structure was only capable of distinguishing between the dynamic multimodal presentation forms used in experiment 2, and not between the durable multimodal presentation forms used in experiment 3. Moreover, in both experiments, it failed to distinguish between the varying magnitudes of multimodal interference for high and low span subjects and thereby to validate the assumed relationship between linguistic complexity and verbal WM capacity in the multimodal domain.

In retrospect, it seems that it was a mistake to bring the parser to its knees using the doubly-embedded structure. Users were unable to successfully parse the complex sentences; when processing this structure, normal parse always failed and processing stalled. In order to continue syntactic parsing, users must have adopted a repair mode, possibly requiring the parse to be restarted. The result was a non-standard processing behaviour. More meaningful answers might have been obtained using a less complex structure with which the parser can cope successfully (e.g., of 4Xint PLUs).

The use of fully predictable syntactic structures

The investigation of the fundamental aspects of multimodal sentence processing was conducted by reductive methods. The investigative studies aimed to provide coherent explanations of multimodal language processing and therefore, constraints of a laboratory setting were devised to maximise experimental control. This is why only two sentence structures were sampled, a single sentence length was used, context was eliminated and intonation and prosody were kept to a minimum. As noted earlier, the use of the doubly-embedded structure in these studies yielded limited information. In contrast, the use of the right-branching structure was capable of demonstrating informative patterns of behaviour. In spite of its fully predictable syntactic structure, subjects tried to comprehend the sentences (allowing pragmatic associations between nouns and verbs to influence their choices). As a result of their natural mode of processing, the right-branching structure was capable of distinguishing between levels of multimodal interference for high and low span subjects in both experiments 2[91] and 3[92]. In order to explain this interference effect, it was suggested that the right-branching sentences placed a high load on low span subjects, urging a further retrieval of the previously processed information. Because resources were low, the recollection of this information was impaired by the concurrent speech as the SAS failed to supervise the coordination of processing between the two information channels.

In the applied study, the investigation of multimodal sentence processing was conducted in a less sterile situation. Task context was introduced and simple priming sentences preceded the test sentences. As noted earlier, these contextual factors had no effect on the fundamental aspects of multimodal sentence processing. However, the inclusion of proper nouns and the added intonation and prosody were suggested to explain the increased use of recall strategies in this study, which made the multimodal interference effect that was identified for low span subjects in the previous studies disappear. As the selected names were arbitrary and free of pragmatic associations, the significance of the pragmatic cues was diminished and subjects processed the fully predictable structures for recall only, focusing on the order of the proper nouns. Intonation and prosody might have supported this retention strategy: the cues that were absent in the previous experiments might have been useful in retaining word order information, thereby complementing the schematic mode of processing in this study.

Including various syntactic structures could encourage processing sentences for comprehension rather than for recall even if proper nouns and intonational cues are used. The reason is that in principle, neither of these factors promotes processing for recall under normal reading/listening conditions. In the absence of additional syntactic structures, a fully predictable long-simple structure with proper nouns, intonation and prosody should not be used to assess processing for comprehension, as responses will involve recall strategies. In this case, a different assessment method of user cost might

---

[91] See combined analysis of experiments 2a and 2b, comprehension rates of low span subjects for long-simple sentences in the dynamic-transient conditions.

[92] See comprehension rates of low span subjects for long-simple sentences in both the dynamic-durable and the static-durable conditions.

be needed: one that promotes a more natural processing of verbal materials. Finally, the processing of short-simple sentence structures might call for a more sensitive method of assessing user cost than those used in this work. The limitations of the methods used are outlined next

Assessment methods of user cost

At the beginning of this research, an attempt was made to use on-line measures of attention and comprehension to answer questions of cognitive compatibility with processing limitations and capabilities of multimodal users. The word-monitoring task was used in experiment 1 to demonstrate the transition from a facilitatory synchronous processing to an interfering asynchronous processing. It intended to show that the accumulation of linguistic demands within long simple and complex sentences imposes varying levels of load on the syntactic and semantic systems and, as a result, slows down the computation of late-position target words by the lexical-access system (more so in complex than in simple sentences). Sentence comprehension served as a complementary task, to ensure that subjects were processing the sentences for comprehension.

The result, however, was poor use of a good dual-task paradigm: the experimental manipulation failed to capture the on-line processing dynamics, particularly as relating to syntactic complexity. The fundamental weakness was the use of the word-monitoring task in conjunction with the task of reading visual text for meaning, using a static-durable presentation of the visual materials. Since the visual materials were available from the onset of the sentence presentation, subjects were able to locate the visual target before it appeared in the auditory channel, which led to a de-sensitisation of the response time measure for the complexity factor. In order to overcome this difficulty, the experiment could have been replicated using three dynamic presentation conditions: a speech-only presentation, a dynamic-transient visual presentation, and a combination of the two. A dynamic presentation of the visual text would have ensured that the visual materials could not be scanned for the target and would have also promoted a more natural processing of sentences presented to both modalities. The reason is that the use of the word-monitoring task in a dynamic multimodal presentation does not necessitate the instruction to attend a specific modality and therefore, the use of catch trials. In this case, the comparison of monitoring times in the multimodal condition with monitoring times in each of the unimodal conditions should be sufficient to reveal which modality facilitates or impairs the processing of the other in multimodal processing. In conclusion, this method might have proved the (unproven) assumption of a transition from a facilitatory synchronous processing to an interfering asynchronous processing under accumulating load conditions.

A different method was used however in subsequent experiments. The possibility that the word-monitoring task may have confounded the experimental results of experiment 1 led to the exclusion of this task from all subsequent experiments. The main assumption was that by giving up the word-monitoring task, one would lose an important source of information regarding the on-line fluctuation of processing resources but in turn would gain an interpretable set of data regarding multimodal sentence processing.

Another option would be to use reading times as well as comprehension rates, in order to regain information about fluctuation of resources consumption in multimodal sentence processing. The eye movement monitoring technique, described in Chapter 2, provides measures of location and duration of eye fixations. It enables the determination of a moment-by-moment profile of processing load across different sentences. It can also provide patterns of eye regressions in a durable presentation of visual materials, thus revealing various processing strategies of the visual text that can be related to the characteristics of the concurrent auditory output. This unobtrusive method can therefore be used for the investigation of fundamental aspects of multimodal processing and also for a further validation of the suggested guidelines. There is no doubt that the use of this technique will provide significantly more information than that obtained using comprehension rates as a single measure of processing cost.

9.5 Future work

The attempt to understand when redundant speech and visual text facilitate the user's processing and comprehension of sentences should be regarded as the beginning of further research. This section will briefly outline possible directions of future work.

As noted in section 9.3.1, this research into fully redundant multimodal presentation can provide the basis for investigating non-redundant multimodal presentation modes, since the principles upon which good coordination of either presentation should be based, are probably similar. The MMDS should be elaborated to account for various non-redundant contents presented by means of multimodality. When multimodality consists of non-redundant information, each presentation channel should be analysed in terms of the linguistic complexity of the verbal content it presents and the memory load placed by its form of presentation. In addition, the relationship between modalities should be made explicit. The relationship between modalities should be specified both at the linguistic level (e.g., referring to the extent of redundancy and complementarity between modalities) and at the presentation level (e.g., referring to the timing of each modality output as this determines the presence of common referents, as well as to the presence of attentional markers such as highlighting and change of font that may vary the strength of association between modalities). The MMUM should be elaborated to include coordination mechanisms of non-redundant multimodality, enabling the prediction of the extent to which one modality increases or decreases the processing cost of the other modality. The expected user cost will depend on the specifications of the contents and their representational forms (for each modality in isolation and for their multimodal combination) by the MMDS. For example, in a partially-redundant multimodality, where visual abbreviations accompany spoken messages (cross-modal priming), the visual channel would be expected to reduce the processing cost of the spoken message if the message is long and the partially redundant visual keywords support a further retrieval of intermediate computational products. Alternatively, the partially redundant visual keywords may direct syntactic processing in their selection. For example, the spoken message may present an object-extracted relative clause: *The reporter who the senator attacked admitted the error*, while the visual keywords include the matrix subject, predicate and object, omitting the embedded clause: *The reporter admitted the error*. On the other hand, in

complementary multimodality, when the meanings conveyed by the two simultaneous outputs are complementary, interference would be expected to occur if one or both channels present demanding content to the user. Furthermore, presenting the visual information dynamically would be expected to increase such interference as dynamic presentation directs the attention of the user and makes the coordination between the two modalities more difficult.

The modelling method used in this research can also contribute to current attempts at investigating combinations of verbal outputs with static or dynamic images. As described in Chapter 1, when verbal information accompanies pictorial information, speech output is superior to visual text output. Consistent with Paivio's (1971) dual-coding theory, users can alternate between the audio and image codes to more effectively obtain required information. In contrast, when the verbal information is delivered by means of visual text output, processing of the visual text competes with processing of the visual images, since they both rely on the same visual resources (Wickens, 1992). The integration of visual images with multimodal presentation of verbal content is informed by inconsistent data (see Chapter 1, section 1.3). The MMDS and the MMUM could be elaborated to account for the inclusion of non-verbal visual contents. The ease of integration between visual images and multimodal verbal information may depend on both the complexity characteristic of the verbal and the non-verbal information and on the memory load imposed by each media. The relationship between the different media should also be accounted for, both at the linguistic level (as specified above) and at the presentation level (e.g., referring to the timing of each media output, to attentional effects and to spatial contiguity aspects of the visual materials). The MMUM should be elaborated to account for coordination mechanisms of verbal and non-verbal information. The expected user cost will depend on the specifications of the contents and their representational forms (for each media in isolation and for their specific combination) by the MMDS.

Finally, a third area for future research would be to test the suggested modelling approach using languages other than English. Both of Gibson's complexity metrics have been tested cross-linguistically to a small degree and were found compatible with complexity judgements in German, Dutch Spanish and Japanese. It would be interesting to find out how the linguistic complexity value of sentences in other languages affects the coordination of verbal information between modalities in various multimodal presentation forms for users with high and low verbal WM capacity.

9.6 Concluding remarks

When this research was started, the question of when and how speech can be combined with visual text to facilitate the user's processing and comprehension of sentences was rather hypothetical. The internet was in its infancy and mobile phones were not a standard accessory. Today the potential for multimodal applications is greater and the question has become extremely relevant. Mobile phones include tiny screens and advances in information compression techniques enable the display of complex multimedia-multimodal presentations by these devices, as well as by home entertainment systems. This makes the reported research exciting and relevant. The design of effective user

interfaces for systems such as these requires a detailed input that can only be provided by an informed research of this type.

Moreover, it seems that the research question was the right question to ask. The combined modelling and empirical approaches that were used in its assessment were successful in providing useful answers, translated to a set of guidelines for effective multimodal presentation of sentences. The pattern of results of this research suggests that multimodal extension may facilitate the user's processing and comprehension of sentences depending on (i) the linguistic complexity of the sentence and (ii) the memory demands incurred by the specific multimodal configuration in relation to those imposed by the primary presentation channel. Specifically, results indicate that a static-durable visual text may reduce the processing cost of demanding sentences delivered by speech. In contrast, speech can be added to various visual forms of text as long as the processing load imposed by the linguistic complexity of the sentence is between low to moderate for a particular individual. Due to methodological problems, the work failed to prove a transition from a facilitatory synchronous processing to an interfering asynchronous processing as linguistic load accumulates. Yet, it succeeded in identifying an effect of speech interference under high load visual text conditions: speech interference was found for all users in comprehending syntactically complex sentences and for low span subjects in comprehending long simple sentences. The realisation of the visual text affected the strength of the speech interference only for complex sentences presented dynamically: interference was larger in the dynamic-durable presentation than in the dynamic-transient presentation. On the other hand, interference was identical for complex sentences presented by means of durable visual text regardless of the dynamism of that text. A common mechanism was used to explain these findings. It was suggested that when processing load increases, processing may breakdown and so users may attempt to retrieve previously processed information. As long as the available resources are sufficient, processing may proceed with no interference. However, when resources are low, the SAS might fail in supervising (i) the competition for resources between the language sub-systems and (ii) the coordination of information between modalities. As a result, the re-activation of intermediate computational products (in a dynamic-transient presentation) or the recollection of visual information (in durable presentations) will be impaired by concurrent speech (regressive eye-movements exacerbate the impairment).

This approach can therefore inform the design process of sentence presentation in redundant multimodal user interfaces and provides the basis for the investigation of non-redundant interfaces, including combinations of verbal outputs with static and dynamic images. However, research could be refined and improved. It should proceed in testing refined assumptions regarding the variation in the user's cognitive cost in response to different contents that are presented in different multimodal configurations. Parallel advances in science can contribute to its future development. Specifically, Gibson's newer complexity metrics can provide detailed predictions of processing cost for low and high span users and these could be assessed for additional levels of linguistic complexity using the eye tracking method. Other than their practical value, results may provide empirical evidence for the validity of this complexity metric, demonstrating the influence of various factors on comprehension performance.

A valedictory note: For some time, GUIs have been the dominant platform for human computer interaction. The GUI-based style of interaction has made computers simpler and easier to use, and the desktop metaphor was successfully used in various applications. However, as the way we use computers changes and computing becomes more pervasive and ubiquitous, GUIs will not easily support the range of interactions necessary to meet users' needs. In order to accommodate a wider range of scenarios, tasks, users and preferences, there is a need to move toward interfaces that are natural, intuitive, adaptive, and unobtrusive. A new metaphor is needed for such interfaces and its nature is yet to be discovered. This metaphor will maximise the use of multiple modalities so the question of how to combine visual and auditory modalities is of increasing significance. Despite the gap between this work and the specification of a fully functional multimedia-multimodal interface, the systematic user-centred approach used in this work advances the achievement of this goal.

References

Anderson, R.C. & Davidson, A. (1988). Conceptual and empirical bases of readability formulas. In A. Davidson and G.M Green (Eds.), *Linguistic complexity and text comprehension.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Baddeley, A.D. (1986). *Working Memory.* Oxford: Oxford University Press.

Baddeley, A.D. (1990). *Human memory: Theory and practice.* Boston, MA: Allyn and Bacon.

Barnard, P. (1985). Interacting cognitive subsystems: A psycholinguistic approach to short-term memory. In A. Ellis (Ed.), *Progress in the psychology of language, Vol. 2* (pp. 197-258). London: Erlbaum.

Barnard, P. & May, J. (1993). Real time blending of data streams: A key problem for the cognitive modelling of user behaviour with multimodal systems. *ESPRIT basic research action 7040, 'AMODEUS' deliverable UM/WP10.*

Bayer, S., Kozierok, R. & Kurtz, J. (1995). Multimodal interfaces on the web. *In Third Python Workshop.* Available on-line: http://www.python.org/workshops/1995-12/papers/bayer/Python-paper.html

Benoit, C., Martin, J., Pelachaud, C., Schomaker, L. & Suhm, B. (1998). Audio-visual and multimodal speech systems. In D. Gibbon (Ed.), *Handbook of Standards and Resources for Spoken Language Systems, Supplement Volume.*

Bernsen, N.O. (1994). A revised generation of the taxonomy of output modalities. *ESPRIT basic research action 7040, 'AMODEUS' deliverable TM/WP11.*

Bernsen, N.O. (2001). Multimodality in language and speech systems: From theory to design support tool. In Granstrom. (Ed.) *Multimodality in Language and Speech Systems.* Kluwer Academic Publishers. Available on-line: http://mate.mip.ou.dk/~nob/publications/MILASS-CHAP-3.10.pdf

Bradley, D.C. & Foster, K.I. (1987). A reader's view of listening. *Cognition, 25,* 103-134.

Brown, G.D.A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavioural Research Methods Instrumentation and Computers, 16* (6), 502-532.

Caplan, D. & Waters, G.S. (1990). Short-term memory and language comprehension: A critical review of the psychological literature. In G. Vallar and T. Shallice (Eds.), *Neuropsychological impairments of short-term memory*. Cambridge: Cambridge University Press, 337-389.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Clark, H.H. & Clark, E.V. (1977). *Psychology and language*. New York: Harcourt Brace.

Connine, C.M., Titone, D. & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 81-94.

Cooper, W.E. & Paccia-Cooper, J.M. (1980). *Syntax and speech*. Cambridge, MA: Harvard University Press.

Coutaz, J. (1992). Multimedia and multimodal user interfaces: A Taxonomy for Software Engineering Research Issues. *In Proceedings of the Second East-Weat HCI conference*, (pp.229-240), St Petersburg, Aug. 1992.

Craig, S.D., Gholson, B. & Driscoll, D.M. (2002). Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology, 94* (2), 428–434.

Cutler, A. (1982). Lexical complexity and sentence processing. In G.B. Flores d'Arcais and R.J. Jarvella (Eds.), *The processes of language understanding*. New York: Wiley.

Dalal, M., Feiner, S., McKeown, K., Pan, S., Zhou, M., Hollerer, T., Shaw, J., Feng, Y. & Fromer, J. (1996). Negotiation for automated generation of temporal multimedia presentations. *Proceedings of the fourth ACM international conference on Multimedia*, (pp.55-64). Boston, Massachusetts. November 18-22, 1996.

Daneman, M. & Carpenter, P.A. (1980). Individual differences in working memory and reading. *Journal of verbal Learning and Verbal Behaviour, 19*, 450-466.

Daneman, M. & Carpenter P.A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory and Cognition, 9*, 561-584.

Dix, A. (1993). Multi-sensory Systems. In A. Dix, J. Finlay, G. Abowd and R. Beale (Eds.), *Human Computer Interaction*. London: Prentice Hall International.

Dowell, J., Life. A. & Salter, I. (1994). The design space for a multimodal multimedia travel facility. *Proceedings of ECCE7, the 7th annual conference of the European Society for Cognitive Ergonomics.*

Dowell, J., Shmueli, Y. & Salter, I. (1995). Applying a cognitive model of the user to the design of a multimodal speech interface. *In Proceedings of the First International Multimodal Interaction workshop, Edinburgh.* July 14-15, 1995.

Eady, J. & Fodor, J.D. (1981). Is centre embedding a source of processing difficulty? *Presented at the Linguistic Society of America annual meeting.*

Eimas, P.D., Hornstein, S.M. & Payton, P. (1990). Attention and the role of the dual code in phoneme monitoring. *Journal of Memory and Language, 29,* 160-180.

Elting, C. (2002). What are multimodalities made of? Modeling output in a multimodal dialogue system. Available on-line: http://www.eml.villabosch.de/english/homes/elting/PUI2001_paper_elting_michelitsch.pdf

Elting, C. & Michelitsch, G. (2001). A multimodal presentation planner for a home entertainment environment. In *Proceedings of the PUI'01 Workshop on Perceptive User Interfaces.*

Faraday, P. & Sutcliffe, A. (1997). Evaluating multimedia presentations. *The New Review of Hypermedia and Multimedia, 3,* 7–37.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language, 30,* 210-233.

Ferreira, F. (1993). The creation of prosody during sentence production. *Psychological Review, 100,* 233-253.

Ferreira, F. & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25,* 348-368.

Ferreira, F., Henderson, J.M., Anes, M.D., Weeks, P.A. & McFarlane, D.K. (1996). Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the auditory moving-window technique. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22* (2), 324-335.

Frazier, L. (1985). Syntactic Complexity. In D. Dowty, L. Karttunen and A. Zwicky (Eds.), *Natural Language Processing Psychological, Computational and Theoretical Perspectives.* Cambridge UK: Cambridge University Press.

Frazier, L. & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14,* 178-210.

Gibson, E.A. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown.* Doctoral dissertation, Carnegie Mellon University. (Available Tech. Rep. CMU-CMT-91-125, from Center for Machine Translation. Carnegie Mellon University, Pittsburgh, PA 15213)

Gibson, E.A. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68,* 1-76.

Gilhooly, K.J. & Logie, R.H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation, 12,* 395-427.

Gould, J.D., Alfaro, L., Barnes, V., Finn, R., Grischkowsky, N. & Minuto, A. (1987). Reading is slower from CRT displays than from paper: Attempts to isolate a single-variable explanation. *Human Factors, 29,* 269-299.

Gould, J.D. & Grischkowsky, N. (1984). Doing the same work with hard copy and with cathode-ray tube (CRT) computer terminals. *Human Factors, 26,* 323-337.

Green, T.R.G. (1991). User modelling: The information processing perspective. In J. Rasmussen, B. Andersen and N.O. Bernsen (Eds.), *Human Computer Interaction.* London: Lawrence Erlbaum Associates.

Greenwald, A.G. (1970). A double stimulation test of ideomotor theory with implications for selective attention. *Journal of Experimental Psychology, 84,* 392-398.

Hanson, V.L. (1981). Processing of written and spoken words: Evidence for common coding. *Memory & Cognition, 9* (1), 93-100.

Hartley, J. (1995). Is this chapter any use? Methods for evaluating text. In J.R. Wilson and E.N. Corlett (Eds.), *Evaluation of human work: A practical ergonomics methodology.* Taylor & Francis.

Hartley, J., Bartlett, S. & Brantwaite, J.A. (1980). Underlining can make a difference-sometimes. *Journal of Educational Research, 73,* 218-224.

Herdman, C.M. (1992). Attentional resource demands of visual word recognition in naming and lexical decisions. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 460-470.

Herdman, C.M. & Dobbs, A.R. (1989). Attentional demands of visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 124-132.

Herfet, T., Kirste T. & Schnaider M. (2001). EMBASSI: Multimodal Assistance for Infotainment and Service Infrastructures. *EC/NSF Workshop Universal on Accessibility of Ubiquitous Computing: Providing for the Elderly,* Alcácer do Sal, Portugal.

Howell, D.C. (1997), *Statistical methods for Psychology (4th edition).* Belmont, CA: Duxbury.

Huls, C. & Bos, E. (1995). Studies into full integration of language and action. *In Proceedings of the International Conference on Cooperative Multimodal Communication (CMC/95),* (pp. 161-174). Eindhoven.

Inhoff, A.W. (1985). The effect of activity on lexical retrieval and postlexical processing during eye fixations in reading. *Journal of Psycholinguistic Research, 14,* 45-56.

Inhoff, A.W. & Rayner, K. (1986). Parafoveal word processing during eye-fixations in reading: Effects of word frequency. *Perception & Psychophysics, 40* (6), 431-439.

Inhoff, A.W., Topolski, R., Vitu, F. & O'Regan, J.K. (1993). Attention demands during reading and the occurrence of brief (express) fixations. *Perception & Psychophysics, 54,* 814-823.

Jarvella, R.J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behaviour, 10,* 409-416.

Juola, J.F., Ward, N.J. & McNamara, T. (1982). Visual search and reading of rapid serial presentations of letter strings, words, and text. *Journal of Experimental Psychology: General, 111,* 208-227.

Just, M.A. & Carpenter, P.A. (1987). *The psychology of reading and language comprehension.* Newton, MA: Allyn & Bacon.

Just, M.A. & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99* (1), 122-149.

Just, M.A., Carpenter, P.A. & Woolley, J. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General, 10,* 833-849.

Kalyuga, S., Chandler, P. & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13,* 351–371.

Kang, T.J. & Muter, P. (1989). Reading dynamically displayed text. *Behaviour & Information Technology, 8,* 33-42.

Kieras, D. & Meyer, D.E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction, 12,* 391-438.

King, J. & Just, M.A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language, 30,* 580-602.

Kintsch, W.A. & Van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-394.

Kirsner, K. & Smith, M.C. (1974). Modality effects in word recognition. *Memory and Cognition, 2,* 637-640.

Kruk, R.S. & Muter, P. (1984). Reading of continuous text on video screens. *Human Factors, 26,* 339-345.

Kryter, K.D. (1972). Speech communications. In H.P. Van Cott and R.G. Kinkade (Eds.), *Human Engineering Guide to System Design.* Washington DC: U.S. Government Printing Office.

Kucera, H. & Francis, W.N. (1967). Computational Analysis of Present-Day American English. Providence: Brown University Press.

Lamel, L., Bennacef, S., Gauvain, J.L., Dartiguest, H. & Temem, J.N. (1998). User Evaluation of the MASK Kiosk. *ICSLP '98,* Sydney.

Landauer, T.K. (1991). Let's get real: A position paper on the role of cognitive psychology in the design of humanly useful and usable systems. In J.M. Carroll (Ed.), *Designing interaction: Psychology at the human-computer interface.* Cambridge University Press.

Lave, J. (1988). *Cognition in Practice: Mind, mathematics, and culture in everyday life.* Cambridge, UK: Cambridge University Press.

Levy, B.A. (1978). Speech processing during reading. In A.M. Lesgold, J.W. Pellegrino, S.D. Fokkema and R. Glaser (Eds.). *Cognitive psychology and instruction* (pp. 123-151). New York: Plenum Press.

Lewis, J.L. (1972). Semantic processing with bisensory stimulation. *Journal of Experimental Psychology, 94,* 455-457.

Lima, S.D. (1987). Morphological analysis in sentence reading. *Journal of Memory & Language, 26,* 84-99.

Luce, P.A., Feustel, T.C. & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors, 25,* 17-32.

Marcus, M.P. (1980). *A theory of syntactic recognition for natural language.* Cambridge, MA: MIT Press.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25,* 71-102.

Marslen-Wilson, W.D. & Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition, 8,* 1-71.

Marslen-Wilson, W.D. & Tyler, L.K. (1987). Against modularity. In J. Garfield (Ed.), *Modularity in knowledge representation and natural-language understanding.* Cambridge, MA: MIT Press.

Martin, R.C., Wogalter, M.S. & Forlano, J.C. (1988). Reading comprehension in the presence of unattended speech and music. *Journal of Memory and Language, 27,* 382-398.

Martin, R.C. (1990). Neuropsychological evidence on the role of short-term memory in sentence processing. In G. Vallar & T. Shallice, *Neuropsychological Impairments of Short-Term Memory.* London: Cambridge University Press.

Martin, R.C. & Feher, E. (1990). The consequences of reduced memory span for the comprehension of semantic versus syntactic information. *Brain and Language, 38,* 1-20.

Masson M.E.J. (1985). Rapid reading processes and skills. In G.E. MacKinnon, and G. Waller, (Eds.), *Reading Research: Advances in Theory and Practice*, Vol 4. Academic Press.

Maybury, M.T. (1993). *Intelligent Multimedia Interfaces.* Cambridge, MA: AAAI Press/MIT Press.

Mayes, J.T. (1994). The 'M-Word': Multimedia interfaces and their role in interactive learning systems. In A.D.N. Edwards and S. Holland (Eds.), *Multimedia Interface Design in Education.* Berlin: Springer-Verlag

McClelland, A. (1999). *Personal Communication*, April 20, 1999.

Meddis, R. (1984). *Statistics Using Ranks: A Unified Approach.* Blackwell Publishers

Miyake, A., Carpenter, P.A. & Just, M.A. (1994). A capacity approach to syntactic comprehension disorders: Making normal adults perform like aphasic patients. *Cognitive Neuropsychology, 11* (6), 671-717.

Moreno, R. & Mayer, R.E. (2000). A coherence effect in multimedia learning: The case for minimizing irrelevant sounds in the design of multimedia instructional messages. *Journal of Educational Psychology, 92,* 117-125.

Muter, P., Latremouille, S.A., Treurniet, W.C. & Beam, P. (1982). Extended reading of continuous text on television screens. *Human Factors, 24,* 501-508.

Muter, P. & Maurutto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited? *Behaviour & Information Technology, 10,* 257-266.

Neal, J.G. & Shapiro, S.C. (1991). Intelligent multi-media interface technology. In J.W. Sullivan and S.W. Tyler (Eds), *Intelligent User Interfaces.* Frontier Series, New York: ACM Press.

Newton, L.D. (1983). The effect of illustrations on the readability of some junior school textbooks. *Reading, 17,* 43-54.

Ni, W., Shankweiler, D. & Crain, S. (1996) Individual differences in working memory and eye-movement patterns in reading relative clause structures. *University of Connecticut Working Papers in Linguistics* 6.

Norman, D.A. & Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In R.J. Davidson, G.E. Schwartz and D. Shapiro (Eds.), *Consciousness and Self-Regulation: Advances in Research and Theory, Vol. 4* (pp. 1-18). New York: Plenum.

Nugent, W.A. (1987). A comparative assessment of compute-based media for presenting job task instructions. *Proceeding of the 31st annual meeting of the Human Factors Society,* (pp. 696-700). Santa Monica: California.

Öquist, G. & Goldstein, M. (2003). Towards an improved readability on mobile devices: Evaluating adaptive rapid serial visual presentation. *Interacting with Computers. Vol. 15* (4), 539-558.

Oviatt, S.L. (1999). Ten myths of multimodal interaction. *Communications of the ACM, 42* (11), 74-81.

Oviatt, S.L. & Cohen, P.R. (2000). Multimodal systems that process what comes naturally. *Communications of the ACM, 43* (3), 45-53.

Paap, K.R. & Noel, R.W. (1991). Dual-route models of print to sound: Still a good horse race. *Psychological Research, 53,* 13-24.

Paivio, A. (1971). Imagery and verbal processes. New York: Holt, Rinehart & Winston.

Pavio, A., Yuille, J.C. & Madigan, S.A. (1968). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology Monograph Supplement, 76* (3, part 2).

Pollastek, A., Rayner, K. & Balota, D.A. (1986). Inferences about eye movement control from the perceptual span in reading. *Perception & Psychophysics, 40,* 123-130.

Posner, M.I., Abdullaev, Y.G., McCandliss, B.D. & Sereno, S.E. (1997). Anatomy, Circuitry, and Plasticity of Reading. In J. Everatt (Ed.), *Visual and attentional processes in reading and dyslexia.* London: Routledge.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology, 7,* 65-81.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior, 22,* 358-374.

Rayner, K. & Duffy, S.A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition, 14,* 191-201.

Rayner, K. & Pollastek, A. (1987). Eye movements in reading: A tutorial review. In M. Coltheart (Ed.), *Attention and Performance XII.* Hove, UK: Lawrence Erlbaum Associates Ltd.

Rayner, K., Well, A.D., Pollastek, A. & Bertera, J.H. (1982). The availability of useful information to the right of fixation in reading. *Perception & Psychophysics, 31,* 537-550.

Reid, D.J., Briggs, N. & Beveridge, M. (1983). The effect of picture upon the readability of a school science topic. *British Journal of Educational Psychology, 53,* 327-335.

Sachs, J.S. (1967). Recognition for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics, 2,* 437-442.

Safran, E.M. & Martin, N. (1990). Short-term memory impairment and sentence processing: A case study. In G. Vallar & T. Shallice (Eds.), *Neuropsychological Impairments of Short-Term Memory.* London: Cambridge University Press.

Seidenberg, M.S., Waters, G.S., Barnes, M.A. & Tanenhaus, M.K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behaviour, 23*, 383-404.

Shallice, T. (1988). *From Neuropsychology to Mental Structure.* London: Cambridge University Press.

Shmueli, Y. (1994). *Modality assignment in multimedia multimodal user interfaces.* Unpublished MSc thesis. University of London.

Shmueli, Y. & Dowell, J. (1999). The multimodal user model: cognitive processes in the integration of visual text and speech. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics, Vol. 4* (pp. 317-327). Ashgate, Aldershot, UK.

Smith, S.L. & Mosier, J.N. (1986). Guidelines for designing user interface software. Report ESD-TR-86-278. Bedford, MA: The MITRE Corporation.

Taylor, W.L. (1953). The cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.

Toglia, M.P. & Battig, W.R. (1978). *Handbook of Semantic Word Norms.* New York: Erlbaum.

Tombaugh, J.W., Arkin, M.D. & Dillon, R.F. (1985). System response factors. *Proceedings of ACM CHI'85 Conference on Human Factors in Computing Systems*, 1-6.

Van Orden, G.C. (1987). A ROWS is A ROSE: Spelling, sound and reading. *Memory & Cognition, 15.*

Vernier, F. & Nigay, L. (2000). A framework for the combination and characterization of output modalities. *Proceedings of DSV-IS2000*, LNCS Springer-Verlag, 32-48.

Ward, M. (2002). *A Template for CALL Programs for Endangered Languages.* MSc Thesis, Dublin City University, Available on-line:
http://www.compapp.dcu.ie/~mward/

Waters, G.S. & Caplan, D. (1996a). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review, 103* (4), 761-772.

Waters, G.S., & Caplan, D. (1996b). Processing resource capacity and the comprehension of garden-path sentences. *Memory and Cognition, 24*, 342-355.

Waters, G.S., Caplan, D. & Hildebrandt, N. (1987). Working memory and written language comprehension. In M. Coltheart (Ed.), *Attention and Performance XII.* Hove, UK: Lawrence Erlbaum Associates Ltd.

Waters, G.S., Caplan, D. & Rochon, E. (1995). Processing resources and sentence comprehension in patients with Alzheimer's disease. *Cognitive Neuropsychology, 12*, 1-30.

Waters, G.S., Komoda, M.K. & Arbuckle, T.Y. (1985). The effects of concurrent tasks on reading. *Journal of Memory & Language, 24*, 27-45.

Waterworth, J.A. & Thomas, C.M. (1985). Why is synthetic speech harder to remember than natural speech? *Proceedings of ACM CHI'85 Conference on Human Factors in Computing Systems,* 201-206.

Wickens, C.D. (1992). *Engineering psychology and human performance.* New York: HarperCollins Publishers Inc.

Wickens, C.D. & Kessel, C. (1980). The processing resource demands of failure detection in dynamic systems. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 564-577.

Wickens, C.D., Sandry, D.L. & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output. *Human Factors*, 25 (2), 227-248.

Wilkinson, R.T. & Robinshaw, H.M. (1987). Proof-reading: VDU and paper text compared for speed, accuracy and fatigue. *Behaviour & Information Technology, 6*, 125-133.

Wright, P. (2001). If Documents Could Talk: Exploring aural and visual language in electronic documents. In L. Degand, Y. Bestgen, W. Spooren and L. Van Waes (Eds.), *Multidisciplinary approaches to discourse.* Amsterdam: Muenster: Stichting Neerlandsteik VU & Nodus Publikationen, 153-161.

W3C Organisation: Web Content Accessibility Guidelines 1.0. Available on-line: http://www.w3.org/TR/WAI-WEBCONTENT/

W3C Organisation: Multimodal Requirements for Voice Mark-up Languages. Work in Progress. Available on-line: http://www.w3.org/TR/2000/WD-multimodal-reqs-20000710

Appendix A

Materials of experiment 1

A.1 Practice

Quantity - Practice (in each presentation condition)

The events that the papers that the man bought covered worry many people. (12)13)
The man bought the papers that covered the events that worry many people. (12)13)
The papers covered the man (F)

Number - Practice (in each presentation condition)
The film that the director that the media likes made won two prizes. (12)13)
The media likes the director that made the film that won two prizes. (12)13)
The director likes the film (F)

Size - practice (catch trial (CT) in each presentation condition)
The boy watched the fireman who extinguished the fire that burnt the big/old house. (13)14)
The fire that the fireman who the boy watched extinguished burnt the big/old house. (13)14)
The boy watched the fireman (T)

A.2 Visual-text

Colour - Text-early
The white cigarette that the match that the man struck lighted exploded. (2)12)
The white man struck the match that lighted the cigarette that exploded. (2)12)
The man struck the match (T)

Size - Text-early
The big plan that the lawyer who the client met devised succeeded. (2)12)
The big client met the lawyer who devised the plan that succeeded. (2)12)
The client devised the plan (F)

Frequency - Text-early
The weekly shipments that the government that the newspaper praised ended helped the guerrillas. (2)14)

The weekly newspaper praised the government that ended the shipments that helped the guerrillas. (2)14)
The government praised the newspaper (F)

Power - Text-early
The strong horse that the man who the lightning killed rode was saved. (2)13)
The strong lightning killed the man who rode the horse that was saved. (2)13)
The horse killed the man (F)

Familiarity - Text-early
The famous race that the horse that the man bought won was shown on TV. (2)15)
The famous man bought the horse that won the race that was shown on TV. (2)15)
The horse won the race (T)

Aesthetics - Text-early
The stylish ship that the waves that the surfers rode beat hit the cliff. (2)14)
The stylish surfers rode the waves that beat the ship that hit the cliff. (2)14)
The surfers hit the cliff (F)

Number - Text-early
The three girls who the boat that the men sold carried ran away from school. (2)15)
The three men sold the boat that carried the girls who ran away from school. (2)15)
The boat carried the girls (T)

Age - Text-early
The old crime that the man who the judge sentenced committed was very infamous. (2)14)
The old judge sentenced the man who committed the crime that was very infamous. (2)14)
The judge sentenced the man (T)

Age - Text-early
The young secretary who the person who the policeman questioned employed disappeared yesterday. (2)13)
The young policeman questioned the person who employed the secretary who disappeared yesterday. (2)13)
The policeman questioned the secretary (F)

Number - CT Text-early
The famous story that the newspaper that the lady prosecuted published was offensive. (2)13)
The famous lady prosecuted the newspaper that published the story that was offensive. (2)13)
The story was offensive (T)

Cost - CT Text-early

The famous mountain that the man who the reporter contacted climbed was covered with snow. (2)15)

The famous reporter contacted the man who climbed the mountain that was covered with snow. (2)15)

The reporter climbed the mountain (F)

Length - Text-late

The rat that the cat that the girl found chased had a short tail. (13)14)

The girl found the cat that chased the rat that had a short tail. (13)14)

The cat chased the rat (T)

Quantity - Text-late

The book that the writer whom the editor liked wrote presents many sensational facts. (12)14)

The editor liked the writer who wrote the book that presents many sensational facts. (12)14)

The writer liked the editor (F)

Distance - Text-late

The shot that the soldier who the mosquito bit fired missed the near target. (13)14)

The mosquito bit the soldier who fired the shot that missed the near target. (13)14)

The mosquito missed the target (F)

Colour - Text-late

The cat that the man who the child saw carried had white legs. (12)13)

The child saw the man who carried the cat that had white legs. (12)13)

The man carried the child (F)

Size - Text-late

The papers that the box that the spy found contained blew into the large room. (14)15)

The spy found the box that contained the papers that blew into the large room. (14)15)

The papers blew into the room (T)

Number - Text-late

The singer that the person who the record company backed signed released four albums. (13)14)

The record company backed the person who signed the singer who released four albums. (13)14)

The record company backed the singer (F)

Number - Text-late

The events that the book that the author published describes occurred fifty years ago. (12)14)

The author published the book that describes the events that occurred fifty years ago. (12)14)

The author published the book (T)

Cost - Text-late
The restaurant that the complex that the businessman owned included served expensive food. (12)13)
The businessman owned the complex that included the restaurant that served expensive food. (12)13)
The complex included the restaurant (T)

Weight - Text-late
The fabrics that the designer who the decorator met created were mostly light and colourful. (13)15)
The decorator met the designer who created the fabrics that were mostly light and colourful. (13)15)
The decorator met the designer (T)

Size - CT Text-late
The letter that the secretary who the manager employed sent was late. (12)12)
The manager employed the secretary who sent the letter that was late. (12)12)
The manager was late (F)

Frequency - CT Text-late
The ball that the pupil who the teacher disliked kicked broke the large window. (13)14)
The teacher disliked the pupil who kicked the ball that broke the large window. (13)14)
The pupil kicked the ball (T)

A.3 Speech

Age - Speech-early
The old banker who the girl who the man married knew died. (2)12)
The old man married the girl who knew the banker who died. (2)12)
The banker married the girl (F)

Aesthetics - Speech-early
The beautiful deer that the man who the boy followed fed was not afraid. (2)14)
The beautiful boy followed the man who fed the deer that was not afraid. (2)14)
The boy fed the deer (F)

Age - Speech-early
The young nurse who the patient who the doctor visited disobeyed works hard. (2)13)
The young doctor visited the patient who disobeyed the nurse who works hard. (2)13)
The patient disobeyed the doctor (F)

Number - Speech-early
The two cases that the judge who the politicians paid influenced had environmental implications. (2)14)
The two politicians paid the judge who influenced the cases that had environmental implications. (2)14)

The politicians paid the judge (T)

Sound - Speech-early
The noisy planes that the sailors who the enemy attacked evaded were bombers. (2)13)
The noisy enemy attacked the sailors who evaded the planes that were bombers. (2)13)
The enemy attacked the planes (F)

Aesthetics - Speech-early
The beautiful lake that the house that the woman bought overlooks is surrounded by trees. (2)15)
The beautiful woman bought the house that overlooks the lake that is surrounded by trees. (2)15)
The house overlooks the lake (T)

Wealth - Speech-early
The rich family that the lawyer who the man confronted represented controlled the food market. (2)15)
The rich man confronted the lawyer who represented the family that controlled the food market. (2)15)
The lawyer represented the family (T)

Size - Speech-early
The small exhibition that the artist who the gallery presented created was extremely expensive. (2)14)
The small gallery presented the artist who created the exhibition that was extremely expensive. (2)14)
The gallery presented the artist (T)

Age - Speech-early
The old car that the man who the lady knew owned crashed. (2)12)
The old lady knew the person who owned the car that crashed. (2)12)
The car crashed (T)

Colour - CT Speech-early
The steel factory that the management that the workers threatened closed was not profitable. (2)14)
The steel workers threatened the management that closed the factory that was not profitable (2)14)
The factory was not profitable (T)

Age - CT Speech-early
The strong article that the reporter who the politician ignored published led to the famous investigation. (2)16)
The strong politician ignored the reporter who published the article that led to the famous investigation. (2)16)
The politician ignored the article (F)

Size - Speech-late
The boats that the rocks that the sea covered sank belonged to the small company. (14)15)
The sea covered the rocks that sank the boats that belonged to the small company. (14)15)
The rocks covered the boats (F)

Number - Speech-late
The actor who the agent who the director met represented performed two days ago. (12)14)
The director met the agent who represented the actor who performed two days ago. (12)14)
The actor met the director (F)

Size - Speech-late
The material that the girl who the designer used cut was not wide enough. (13)14)
The designer used the girl who cut the material that was not wide enough. (13)14)
The designer used the girl (T)

Quantity - Speech-late
The strategies that the general who the soldiers respected used led to many victories. (13)14)
The soldiers respected the general who used the strategies that led to many victories. (13)14)
The general used the strategies (T)

Size - Speech-late
The window that the ball that the boy threw hit broke into little pieces. (13)14)
The boy threw the ball that hit the window that broke into little pieces. (13)14)
The ball hit the window (T)

Size - Speech-late
The project that the architect who the contractor met planned received huge recognition. (12)13)
The contractor met the architect who planned the project that received huge recognition. (12)13)
The contractor planned the project (F)

Number - Speech-late
The dog that the boy that the woman hugged missed disappeared two weeks ago. (12)14)
The woman hugged the boy who missed the dog that disappeared two weeks ago. (12)14)
The dog missed the boy (F)

Power - Speech-late
The article that the journalist who the readers liked wrote contained strong language. (12)13)
The readers liked the journalist who wrote the article that contained strong language. (12)13)
The readers liked the article (F)

Number - Speech-late
The man who the woman who the lawyer met married was convicted five days ago. (13)15)

The lawyer met the woman who married the man who was convicted five days ago. (13)15)
The lawyer met the woman (T)

Sound - CT Speech-late
The doctor who the villagers who the chapter describes killed lived in a poor district. (14)15)
The chapter describes the villagers who killed the doctor who lived in a poor district. (14)15)
The chapter describes the villagers (T)

Number - CT Speech-late
The student that the lecturer who the committee reproached blamed did not steal the expensive equipment. (15)16)
The committee reproached the lecturer who blamed the student that did not steal the expensive equipment. (15)16)
The student did not steal the equipment (T)

A.4 Multimodal Visual-Attend (MMVA)

Colour - MMVA-early
The black car that the man who the dog bit drove crashed. (2)12)
The black dog bit the man who drove the car that crashed. (2)12)
The dog bit the man (T)

Cost - MMVA-early
The expensive video that the artist that the producer met made won the film prize. (2)15)
The expensive producer met the artist who made the video that won the film prize. (2)15)
The producer made the artist (F)

Size - MMVA-early
The small car that the person who the hospital admitted drove crashed. (2)12)
The small hospital admitted the person who drove the car that crashed. (2)12)
The person drove the car (T)

Colour - MMVA-early
The white castle that the hill that the snow covered overlooked was very impressive. (2)14)
The white snow covered the hill that overlooked the castle that was very impressive. (2)14)
The castle overlooked the hill (T)

Height - MMVA-early
The tall post that the ball that the player kicked hit blocked the goal. (2)14)
The tall player kicked the ball that hit the post that blocked the goal. (2)14)
The player kicked the post (F)

Age - MMVA-early
The young baby who the woman who the waiter served stroked was crying. (2)13)
The young waiter served the woman who stroked the baby who was crying. (2)13)
The waiter served the woman (T)

Frequency - MMVA-early
The monthly column that the reporter who the magazine fired wrote was controversial. (2)13)
The monthly magazine fired the reporter who wrote the column that was controversial. (2)13)
The reporter wrote the column (T)

Age - MMVA-early
The old house that the woman who the man liked bought was very expensive (2)14)
The old man liked the woman who bought the house that was very expensive. (2)14)
The woman liked the man (F)

Weight - MMVA-early
The heavy groceries that the woman who the man pushed dropped fell on the road. (2)15)
The heavy man pushed the woman who dropped the groceries that fell on the road. (2)15)
The woman pushed the groceries (F)

Aesthetics - CT MMVA-early
The new/elegant computer that the company that the designer joined developed performed best. (2)13)
The new/elegant designer joined the company that developed the computer that performed best. (2)13)
The computer performed best (T)

Aesthetics - CT MMVA-early
The young/beautiful girl who the doctor who the nurse admired cured was dying. (2)13)
The young/beautiful nurse admired the doctor who cured the girl who was dying. (2)13)
The nurse was dying (F)

Weight - MMVA-late
The man who the woman who the dog bit likes dropped the heavy box. (13)14)
The dog bit the woman who likes the man who dropped the heavy box. (13)14)
The dog bit the woman (T)

Number - MMVA-late
The lady who the man who the police chased killed was fifty years old. (12)14)
The police chased the man who killed the lady who was fifty years old. (12)14)
The lady was fifty years old (T)

Number - MMVA-late
The tiger that the men who the master paid caught killed five dogs. (12)13)
The master paid the men who caught the tiger that killed five dogs. (12)13)
The tiger caught the master (F)

Power - MMVA-late
The study that the scientist who the committee appointed published had a strong impact. (13)14)
The committee appointed the scientist who published the study that had a strong impact. (13)14)
The committee published the study (F)

Size - MMVA-late
The song that the band that the audience clapped played was a big hit. (13)14)
The audience clapped the band that played the song that was a big hit. (13)14)
The band played the song (T)

Number - MMVA-late
The allegations that the judge who the reporter interviewed dismissed were made two days ago. (13)15)
The reporter interviewed the judge who dismissed the allegations that were made two days ago. (13)15)
The reporter dismissed the judge (F)

Number - MMVA-late
The banker who the man who the police arrested murdered was sixty years old. (12)14)
The police arrested the man who murdered the banker who was sixty years old. (12)14)
The man murdered the banker (T)

Wealth - MMVA-late
The doctor who the man who the police arrested injured worked in the poor neighbourhood. (14)15)
The police arrested the man who injured the doctor who worked in the poor neighbourhood. (14)15)
The police arrested the man (T)

Weight - CT MMVA-late
The jewellery that the man who the fireman saved stole was in the steel/heavy safe. (14)15)
The fireman saved the robber who stole the jewellery that was in the steel/heavy safe. (14)15)
The journalist stole the jewellery (F)

Cost - CT MMVA-late
The merchandise that the salesman who the boy assisted sold was relatively old/cheap. (13)13)
The boy assisted the salesman who sold the merchandise that was relatively old/cheap. (13)13)
The salesman sold the merchandise (T)

A.5 Multimodal Auditory Attend (MMAA)

Age - MMAA-early
The young woman who the man who the detective questioned married disappeared last week. (2)14)
The young detective questioned the man who married the woman who disappeared last week. (2)14)
The woman married the detective (F)

Age - MMAA-early
The young couple who the man who the boy saw robbed lived next door. (2)14)
The young boy saw the man who robbed the couple who lived next door. (2)14)
The man saw the couple (F)

Size - MMAA-early
The huge house that the trees that the storm uprooted hit was saved. (2)13)
The huge storm uprooted the trees that hit the house that was saved. (2)13)
The house was saved (T)

Frequency - MMAA-early
The daily column that the journalist who the newspaper employed wrote was very successful. (2)14)
The daily newspaper employed the journalist who wrote the column that was very successful. (2)14)
The journalist wrote the column (T)

Wealth - MMAA-early
The rich company that the accountant who the manager sued cheated went bankrupt. (2)13)
The rich manager sued the accountant who cheated the company that went bankrupt. (2)13)
The manager cheated the company (F)

Wealth - MMAA-early
The rich county that the lawyer who the farmer challenged represented possessed the land. (2)15)
The rich farmer challenged the lawyer who represented the county that possessed the land. (2)15)
The farmer challenged the lawyer (T)

Aesthetics - MMAA-early
The gorgeous garden that the flat that the actor bought overlooked bloomed. (2)12)
The gorgeous actor bought the flat that overlooked the garden that bloomed. (2)12)
The flat overlooked the garden (T)

Cost - MMAA-early
The cheap flat that the person who the estate agent liked sold was extremely small. (2)15)
The cheap estate agent liked the person who sold the flat that was extremely small. (2)15)
The estate agent liked the person (T)

Weight - MMAA-early
The heavy plate that the waiter who the guest pushed dropped broke. (2)12)
The heavy guest pushed the waiter who dropped the plate that broke. (2)12)
The waiter pushed the guest (F)

Familiarity - CT MMAA-early
The young/famous man who the lawyer who the journalist interviewed defended was innocent. (2)13)
The young/famous journalist interviewed the lawyer who defended the man who was innocent. (2)13)
The journalist interviewed the man (F)

Wealth - CT MMAA-early
The famous/rich banker who the lawyer who the couple met represented went bankrupt. (2)13)
The famous/rich couple met the lawyer who represented the banker who went bankrupt. (2)13
The couple met the lawyer (T)

Size - MMAA-late
The stock that the broker who the man respected purchased made a huge profit. (13)14)
The man respected the broker who purchased the stock that made a huge profit. (13)14)
The man purchased the stock (F)

Quantity - MMAA-late
The museum that the architect who the lecturer praised planned housed many great artists. (12)14)
The lecturer praised the architect who planned the museum that housed many great artists. (12)14)
The lecturer planned the museum (F)

Colour - MMAA-late
The waiter who the comedian who the couple watched amused opened the red wine. (13)14)
The couple watched the comedian who amused the waiter who opened the red wine. (13)14)
The comedian amused the waiter (T)

Colour - MMAA-late
The man who the woman who the cat liked married eats red meat. (12)13)
The cat liked the woman who married the man who eats red meat. (12)13)
The man liked the cat (F)

Length - MMAA-late
The girl who the ball that the boy kicked hit had long hair. (12)13)
The boy kicked the ball that hit the girl that had long hair. (12)13)
The girl hit the boy (F)

Colour - MMAA-late

The car that the man that the bee stung drove had a white roof. (13)14)

The bee stung the man who drove the car that had a white roof. (13)14)

The bee stung the man **(T)**

Age - MMAA-late

The jewellery that the boy who the police chased stole belonged to the old lady. (14)15)

The policeman chased the boy who stole the jewellery that belonged to the old lady. (14)15)

The boy stole the jewellery **(T)**

Size - MMAA-late

The library that the architect who the contractor praised planned was large and spacious. (12)14)

The contractor praised the architect who planned the library that was large and spacious. (12)14)

The contractor praised the architect **(T)**

Number - MMAA-late

The flat that the woman who the architect met bought was built two years ago. (13)15)

The architect met the woman who bought the flat that was built two years ago. (13)15)

The architect bought the flat (F)

Quantity - CT MMAA-late

The company that the person who the doctor examined managed lost important/many clients last year. (12)15)

The doctor examined the person who managed the company that lost important/many clients last year. (12)15)

The company lost clients last year **(T)**

Length - CT MMAA-late

The letters that the drawer that the girl opened contained were held together by a black/long ribbon. (16)17)

The girl opened the drawer that contained the letters that were held together by a black/long ribbon. (16)17)

The drawer contained the girl (F)

Appendix B

An analysis of the monitoring responses in experiment 1

Table B.1 presents the distribution of responses observed for the monitoring task in the unimodal and the multimodal conditions in terms of the signal detection theory.

Table B.1
Distribution of Responses in the Unimodal and the Multimodal Conditions (%): By Signal Detection Categories[93]

Multimodality	Signal Detection Categories					
	Hit ("2" in the presence of a catch trial)	Miss ("0" in the presence of a catch trial)	Missing responses for catch trials	Correct Rejection ("0" in the absence of a catch trial)	False Alarm ("2" in the absence of a catch trial)	Missing responses for regular trials
Unimodal	85%	11%	4%	91%	7%	2%
Multimodal	65%	32%	3%	95%	3%	2%

Due to the small number of catch trials, the data was analysed using non-parametric tests. Missing responses were excluded from all analyses. The Wilcoxon Signed Ranks test yielded a main effect of multimodality for the Hit category ($Z = -2.975$; $p < .01$). This effect suggests that subjects in the multimodal conditions must have allocated attention to the unattended channel. The table also demonstrates the different patterns of performance in the unimodal and the multimodal conditions. In the unimodal conditions, subjects exhibited a balanced performance: the high proportion of correctly capturing the catch trials (Hit) equals the proportion of correctly capturing the target words (Correct Rejection). On the other hand, subjects in the multimodal conditions were highly successful in correctly capturing the target words (Correct Rejection) but relatively poor in responding correctly to present catch trials (Hit).

It is suggested that the catch trials in the unimodal and the multimodal conditions are not entirely comparable. For example, the position variable is meaningless in the unimodal conditions, since subjects had to process a complete catch trial sentence in order to decide that it did not contain a compatible adjective. A further examination of the Hit data was thus conducted separately for these

[93] For a full explanation of the Signal Detection Categories, see Chapter 5, section 5.3.5.

conditions. The examination of the data consisted of conducting multiple comparisons among pairs of means created by the experimental variables. The Wilcoxon Signed Ranks test was applied to examine the effects that involve solely the within-subjects variables, and the Mann-Whitney test was used to examine all effects that include the complexity variable.

For the unimodal conditions, the analyses produced neither a main effect of modality (Z = -.882), nor a main effect of complexity (Mann-Whitney U = 51.00; Wilcoxon W = 129.00; Z = -.973). Also, the interaction between modality and complexity did not reach significance (Mann-Whitney U = 55.00; Wilcoxon W = 133.00; Z = -.721); (see Table B.2). Similarly, the analyses conducted for the multimodal conditions did not yield a main effect of modality (Z = -.269), a main effect of complexity (Mann-Whitney U = 64.50; Wilcoxon W = 130.50; Z = -.094), a main effect of position (Z = -.294), an interaction between modality and complexity (Mann-Whitney U = 63.00; Wilcoxon W = 129.00; Z = -.190), an interaction between complexity and position (Mann-Whitney U = 43.00; Wilcoxon W = 109.00; Z = -1.450), an interaction between modality and position (Z = -.824) or a triple interaction between complexity, position and modality (Mann-Whitney U = 53.00; Wilcoxon W = 119.00; Z = -.827); (see Table B.3).

Table B.2

Mean Recognition Rate of Catch Trial Sentences for the Unimodal Conditions (%): By Complexity and Modality (Standard Errors in Parentheses)

	Modality-based Conditions	
Complexity Level	Visual	Speech
Simple	86%	86%
	(6%)	(7%)
Complex	79%	90%
	(6%)	(7%)

Table B.3

Mean Recognition Rate of Catch Trial Sentences for the Multimodal Conditions (%): By Complexity, Modality and Word-Position (Standard Errors in Parentheses)

	Presentation Conditions by Position			
Complexity Level	MMVA-Early	MMVA-Late	MMAA-Early	MMAA-Late
Simple	55%	73%	59%	77%
	(10%)	(9%)	(12%)	(10%)
Complex	75%	54%	63%	67%
	(10%)	(9%)	(12%)	(10%)

The pattern of results in the multimodal conditions suggests that the manipulation of attention in this study was at least partially unsuccessful. Subjects faced difficulties in ignoring the targets when they appeared in the unattended channel. Since speech cannot be ignored, this finding is not surprising for

the visual-based conditions. However, the absence of a main effect of modality in the multimodal conditions indicates that the visual text was not ignored in the MMAA conditions.

A separate analysis was conducted for the Correct Rejection category. The Correct Rejection category refers to the recognition rate of target words in the regular trials (Correct Rejection = 100% - (False Alarm + missing responses[94] for regular trials)). The motivation of this analysis was to find out whether the word-monitoring task was equal in ease in all conditions. Any differences in ease of response would reflect on the overall pattern of results reported throughout this section. It was decided to analyse the Correct Rejection data using non-parametric tests[95]. This examination consisted of conducting multiple comparisons among pairs of means created by the experimental variables. The Wilcoxon Signed Ranks test was applied to examine the effects that involve solely the within-subjects variables, and the Mann-Whitney test was used to examine all effects that include the complexity variable.

The analyses yielded a main effect of position: the data indicates a higher Correct Rejection rate for early-position targets (95%) than for late-position ones (91%) ($Z = -3.142$; $p < .01$). This might imply that the short presentation time provided for each sentence made it difficult to perform the word-monitoring task for late-position targets. Moreover, the analyses yielded a significant effect of multimodality ($Z = -3.207$; $p < .01$): the data indicates a higher Correct Rejection rate in the multimodal conditions (95%) relative to the unimodal conditions (91%). This implies that in both modality-based conditions, regardless of sentence complexity, it was more difficult to correctly capture target words for regular trials in the unimodal conditions relative to the multimodal conditions. Moreover, the view of difficulties in the unimodal conditions gains support from an inspection of subjects' data. This inspection reveals that in each of the unimodal conditions, there were two target words which some third of all subjects mistook for catch trials. The rate of mistakes does not appear to be affected by the complexity of the presented sentences. No indication of specific difficulties was found for the multimodal conditions. It is thus proposed that the word-monitoring task was more difficult in the unimodal conditions relative to the multimodal conditions.

All other analyses did not yield significant effects. They did not yield a main effect of complexity (Mann-Whitney $U = 56.50$; Wilcoxon $W = 134.50$; $Z = -.595$) or a main effect of modality ($Z = -.298$). Moreover, the following two-way interactions were not significant: the interaction between complexity and multimodality (Mann-Whitney $U = 52.00$; Wilcoxon $W = 118.00$; $Z = -.915$), the interaction between modality and complexity (Mann-Whitney $U = 61.50$; Wilcoxon $W = 127.50$; $Z = -.281$), the interaction between position and complexity (Mann-Whitney $U = 59.50$; Wilcoxon $W = 137.50$; $Z = -.416$), the interaction between modality and position ($Z = -1.110$) and the interaction between multimodality and position ($Z = -.343$).

---

[94] Missing responses are implicitly considered as False alarm responses in this analysis (i.e., pressing "2" in the absence of a catch trial)

[95] The exploration of the Correct Rejection data revealed that the assumption of normality was not met for most sub-conditions created by the complexity, modality, multimodality and position variables. The assumption of homogeneity of variance was met for all conditions bar the speech-early and the MMAA-early conditions.

The interaction between modality and multimodality approached significance ($Z = -1.875$; $p < .07$) (see Table B.4). However, both simple main effects failed to reach significance ($Z_{unimodal} = -1.027$, $Z_{Multimodal} = -1.489$).

Table B.4
Mean Correct Rejection Rate in Regular Trials (%): By Modality and Multimodality
(Standard Errors in Parentheses)

	Modality-based Conditions	
Multimodality	Visual-based	Auditory-based
Unimodal	90%	92%
	(1%)	(2%)
Multimodal	96%	94%
	(1%)	(1%)

All triple interactions were not significant: the interaction between complexity, modality and multimodality (Mann-Whitney U = 46.50; Wilcoxon W = 124.50; Z = -1.246), the interaction between complexity, position and multimodality (Mann-Whitney U = 37.50; Wilcoxon W = 115.50; Z = -1.808), the interaction between complexity, position and modality (Mann-Whitney U = 55.00; Wilcoxon W = 121.00; Z = -.692) and the interaction between modality, multimodality and position (Z = -.854). Finally, the four-way interaction did not reach significance (Mann-Whitney U = 53.50; Wilcoxon W = 119.50; Z = -.798).

Appendix C

Non-parametric tests in experiment 2a

Due to the effect of the pragmatic complexity variable on comprehension rates, only small number of sentences was selected for each presentation condition in the separate analyses conducted for the single-line and the double-line data of experiment 2a. As a result o the small samples, there were serious departures from the normality assumption required for a valid parametric test. A different - non-parametric procedure was needed, one that calls for less stringent assumptions about the data. Since the commonly used non-parametric tests are suitable for simple one-way layout designs, the selected approach involved combining two existing tests. The Mann-Whitney test was selected to examine all contrasts involving the between subjects variable and the Wilcoxon matched-pairs signed-rank test was selected to examine all contrasts involving the within subjects variables (Alastair McClelland, April 20, 1999, personal communication, see also Meddis, 1984).

In each of the separate analyses conducted for the single-line and the double-line data, performance rates (comprehension rates and response times) were collected separately for each subject in each presentation condition. Four measures were collected for each subject:
Dynamic-durable visual
Dynamic-durable multimodal
Dynamic-transient visual
Dynamic-transient multimodal

Contrasts involved various computations based on these rates:
1. Mean performance rate = (Dynamic-durable visual + Dynamic-durable multimodal + Dynamic-transient visual + Dynamic-transient multimodal)/4
2. Durable performance rate = (Dynamic-durable visual + Dynamic-durable multimodal)/2
3. Transient performance rate = (Dynamic-transient visual + Dynamic-transient multimodal)/2
4. Durability difference = Durable performance rate - Transient performance rate
5. Unimodal performance rate = (Dynamic-durable visual + Dynamic-transient visual)/2
6. Multimodal performance rate = (Dynamic-durable multimodal + Dynamic-transient multimodal)/2
7. Multimodality difference = Unimodal performance rate - Multimodal performance rate
8. Multimodality difference for durable conditions = Dynamic-durable visual - Dynamic-durable multimodal
9. Multimodality difference for transient conditions = Dynamic-transient visual - Dynamic-transient multimodal
10. Multimodality difference for different durability conditions = Multimodality difference for durable conditions - Multimodality difference for transient conditions

The experimental effects were investigated as described next:

- Main effect of complexity: The Mann-Whitney test was used to test the prediction that performance rates in the complex condition are lower than performance rates in the simple condition. This involved ranking the combined "Mean performance rate" values from lowest to highest without regard to their complexity condition membership and then comparing the sum of ranks assigned to each complexity condition.
- Main effect of durability: The Wilcoxon matched-pairs signed-ranks test was used to test whether the "Durability difference" between the "Durable performance rate" and the "Transient performance rate" is significant (see contrast number 4). This involved ranking the "Durability difference" scores from lowest to highest without regard to the sign of the differences and then assigning the algebraic sign of the differences to the signs themselves. Positive and negative ranks were summed separately and the difference of their means was assessed statistically.
- Interaction between durability and complexity: The Mann-Whitney test was used to test whether this "Durability difference" is affected by sentence complexity. This involved ranking the "Durability difference" values from lowest to highest without regard to their complexity condition membership and then comparing the sum of ranks assigned to each complexity condition. In case that this interaction was significant, simple main effects of durability were assessed for each level of complexity using the Wilcoxon matched-pairs signed-ranks test.
- Main effect of multimodality: The Wilcoxon matched-pairs signed-ranks test was used to test whether the "Multimodality difference" between the "Unimodal performance rate" and the "Multimodal performance rate" is significant (see contrast number 7). This involved ranking the "Multimodality difference" scores from lowest to highest without regard to the sign of the differences and then assigning the algebraic sign of the differences to the signs themselves. Positive and negative ranks were summed separately and the difference of their means was assessed statistically.
- Interaction between multimodality and complexity: The Mann-Whitney test was used to test whether this "Multimodality difference" is affected by sentence complexity. This involved ranking the "Multimodality difference" values from lowest to highest without regard to their complexity condition membership and then comparing the sum of ranks assigned to each complexity condition. In case that this interaction was significant, simple main effects of multimodality were assessed for each level of complexity using the Wilcoxon matched-pairs signed-ranks test.
- Interaction between durability and multimodality: The Wilcoxon matched-pairs signed-ranks test was used to test whether the "Multimodality difference for durable conditions" differs from the "Multimodality difference for transient conditions" (see contrast number 10). This involved ranking the "Multimodality difference for different durability conditions" scores from lowest to highest without regard to the sign of the differences and then assigning the algebraic sign of the differences to the signs themselves. Positive and negative ranks were summed separately and the difference of their means was assessed statistically. In case that this interaction was significant, simple main effects of multimodality were assessed for each level of durability using the Wilcoxon matched-pairs signed-ranks test.

- Interaction between durability, multimodality and complexity: The Mann-Whitney test was used to test whether this "Multimodality difference for different durability conditions" is affected by sentence complexity. This involved ranking the "Multimodality difference for different durability conditions" values from lowest to highest without regard to their complexity condition membership and then comparing the sum of ranks assigned to each complexity condition. In case that this triple interaction was significant, two-ways interactions between durability and multimodality were assessed for each level of complexity using the Wilcoxon matched-pairs signed-ranks test. This involved ranking the "Multimodality difference for different durability conditions" scores for each level of complexity from lowest to highest without regard to the sign of the differences and then assigning the algebraic sign of the differences to the signs themselves. Positive and negative ranks were summed separately and the difference of their means was assessed statistically. Finally, in case that these two-ways interactions reached significance, simple main effects of multimodality were assessed for each level of durability in each complexity condition using the Wilcoxon matched-pairs signed-ranks test.

Appendix D

Materials of the applied study

D.1 Practice: Journalism Scenario

D.1.1 Practice: Journalism Scenario – Visual Text Presentation

- An exclusive

We might have an exclusive. (5)
Paul Stuart met Julian Herbert who is a good source for celebrity gossip. (13)
A1B Did Paul Stuart meet Julian Herbert? (Yes)

- Details

There are still details unfolding. (5)
Tracey Beard spoke with Marie Garret who uncovered the story that became front-page news. (14)
B3D Did Marie Garret become front-page news? (No)

- A thorough article

This was a very thorough article. (6)
Glenn Culley praised Thomas Wood who researched the subject. (9)
B1A Did Thomas Wood praise Glenn Culley? (No)

D.1.2 Practice: Journalism Scenario – Speech Presentation

- Another source

We have to find another source. (6)
Nigel Baker offended Andrew Saunders who is unwilling to discuss the article that covered the scandal. (16)
A3D Did Nigel Baker cover the scandal? (No)

- One side

We have heard only one side. (6)
Jessica Walsh talked to Ann Mellor who is the central figure of the story. (14)
B2C Is Ann Mellor the central figure of the story? (Yes)

- Government affairs

We need an analyst for government affairs. (7)
James Green approved of Richard Hartley who published a book that deals with modern politics. (15)
A1B Did James Green approve of Richard Hartley? (Yes)

D.1.3 Practice: Journalism Scenario – Multimodal Presentation

- Reporters

The reporters need to be more aggressive. (7)
Darren Philips avoided Adam Walsh who has been difficult in the past. (12)
A1B Did Darren Philips avoided Adam Walsh? (Yes)

- Personal details

Some personal details are starting to come through. (8)
Diana Parker encountered Melanie Hammond who knows the MP that lied in court. (13)
A3D Did Diana Parker lie in court? (No)

- Another author

I would like another author for this report. (8)
Lucy Palmer admires Katherine Garfield who used to write for The Times. (12)
B2C Did Katherine Garfield use to write for The Times? (Yes)

D.2 Designers Recruitment Agency Scenario – Visual Text Presentation

D.2.1 Designers Recruitment Agency Scenario - 2 Clauses

- Graphics people

We're still looking for graphics people. (7)
Mark Farnham called James White who has recently graduated from the Royal College of Art. (15)
A2C Has Mark Farnham recently graduated from the Royal College of Art? (No)

- Our banking system

There seem to be new problems with our banking system. (11)
Sara Weymouth has criticised Karen Brewer who works in the accounts department. (12)
A1B Has Sara Weymouth criticised Karen Brewer? (Yes)

- Short list

We are down to the short list. (8)
Jane Alexander has examined Megan Brown who applied for the systems design job. (13)
B1A Has Megan Brown examined Jane Alexander? (No)

- The invoicing process

The invoicing process is not working properly. (8)
Rebecca Jones has blamed Tara Mitchell who delayed the payment for the opera project. (14)
B2C Did Tara Mitchell delay the payment for the opera project? (Yes)

- Inexperienced candidates

There are plenty of inexperienced candidates. (7)
Lisa Brown has contacted Sarah Nolan who has just started at Middlesex University. (13)
C1B Did Middlesex University contact Sarah Nolan? (No)

- The web design project

Don't forget the web design project. (7)
Guy Owens has replied to Philip Kelly who works in the internet department. (13)
A1B Has Guy Owens replied to Philip Kelly? (Yes)

- People with good fashion sense

It is difficult to find people with good fashion sense. (11)
Alice Matthews has faxed Sandra Whittle who manages the styling department at Selfridges. (13)
B1C Has Sandra Whittle faxed the styling department at Selfridges? (No)

- Attendance

You don't need to attend. (5)
Stuart Jones has emailed James Minister who will conduct the interview. (11)
B2C Will James Minister conduct the interview? (Yes)

D.2.2 Designers Recruitment Agency Scenario - 3 Clauses

- The country inn project

The country inn project is in hand. (8)
Suzan Archer has interviewed Angela Roberts who is currently working for Future Interiors Ltd that designed the Four Seasons hotel. (20)
A1B Has Suzan Archer interviewed Angela Roberts? (Yes)

- Head hunting

Our head hunting for the chief executive seems to be going well. (13)
Alice Russell has approached Niki Carlson who runs the corporate department that is currently working on the Harrods account. (19)
A2C Does Alice Russell run the corporate department? (No)

- A feature editor

We still need a feature editor. (7)
Brian Butler has mentioned Sam Davies who used to write for Wallpaper magazine that promotes Architectural European projects. (18)
B2C Did Sam Davies use to write for Wallpaper magazine? (Yes)

- Back door

Good applicants are coming through the back door. (9)
Sam Wilson has written to Justine Clark who works for the company that is looking for an experienced manager. (19)
B2A Does Justine Clark work for Sam Wilson? (No)

- Replying stage

We are replying to all the people in the short list. (11)
Peter Brook has answered Neil Knowles who used to write for Living Etc. magazine that focuses on domestic decoration projects. (20)
A1B Has Peter Brook answered Neil Knowles? (Yes)

- Our on-line team

Our on-line team is doing well. (6)
Paula White has thanked Chloe Patterson who directs the internet department that recruited the experienced software designers. (17)
C1A Has the internet department thanked Paula White? (No)

- Display design people

We are still short of people with experience of display design. (11)
Anna Walker has recommended Natalie Mitchell who is currently working for
Wentworth Design that specialises in innovative retail projects. (19)
B2C Is Natalie Mitchell currently working for Wentworth Design? (Yes)

- The front cover job

We've found the illustrator for the front cover job. (9)
James Reid has impressed Tim Pierce who works in the media department that specialises in graphic designers. (17)
B1C Has Tim Pierce impressed the media department? (No)

D.2.3 Designers Recruitment Agency Scenario - Filler Messages

- Untested managers

We don't want any more untested managers. (7)
Robert Murray represents Dan Bishop who supervised the South Bank project that fell apart. (14)
C3D Did the South Bank project that fall apart? (Yes)

- Being careful

We need to be more careful. (6)
James Michael confronted George Edwards who recommended the designer who stole the goods. (13)
B3D Did George Edwards steal the goods? (No)

- You should clear the air between your colleagues. (9)

Helen Grant reproached Alma Gordon who accepted the client that delayed the instalments. (13)
A3D Did Helen Grant delay the instalments? (No)

D.3 Film Scenario – Speech Presentation

D.3.1 Film Scenario - 2 Clauses

- An understudy

We're still looking for an understudy. (6)
Jeff Barnes called Greg Powell who has recently graduated from London Drama School. (13)
B2C Has Greg Powell recently graduated from London Drama School? (Yes)

- The rehearsals

The rehearsals are tense. (4)
Claire Barker has criticised Shelly Gilbert who plays the female lead. (11)
C1B Has the female lead criticised Shelly Gilbert? (No)

- TV people

We've had some interest from TV people. (7)
Chris Harris has examined Barry Turner who applied for the producer job. (12)
A1B Has Chris Harris examined Barry Turner? (Yes)

- Keeping to schedule

The casting department have not kept to schedule. (8)
Liz Bell has blamed Jackie Cole who delayed the rehearsals for the new scene. (14)
A2C Did Liz Bell delay the rehearsals for the new scene? (No)

- Junior assistants

Junior assistants will not be a problem. (7)
Louise Francis has contacted Anne Godwin who has just started at Goldsmith College. (13)
B2C Has Anne Godwin just started at Goldsmith College? (Yes)

- The kitchen scene

We don't have enough footage of the kitchen scene. (9)
Neil Marshall has replied to Simon Green who works in the editing department. (13)
B1C Has Simon Green replied to the editing department? (No)

- Theatre experience

We need someone with theatre experience. (6)
Martin Clarke has faxed John Pearson who manages the lighting department at the Royal Shakespeare Company. (16)
B2A Does John Pearson manage Martin Clarke? (No)

- Showing up

You don't have to show up. (6)
Patricia Gordon has emailed Carol Johns who will conduct the final audition. (12)
A1B Has Patricia Gordon emailed Carol Johns? (Yes)

D.3.2 Film Scenario - 3 Clauses

- The daughter role

We've found an ideal person for the daughter role. (9)
Sylvia Holden has interviewed Julia Gray who is currently acting in the West End play that won the Laurence Olivier Award. (21)
B2C Is Julia Gray currently acting in the West End play? (Yes)

- Wardrobe assistant

We need one more wardrobe assistant. (6)
Anne Lewis has approached Mary Lawrence who works in the clothing department that is currently creating costumes for the opening scene. (21)
A2C Does Anne Lewis work in the clothing department? (No)

- Unknowns for the lead

We are also looking at unknowns for the lead. (9)
Jonathan Knight has mentioned Dan Shields who used to act in a local theatre group that performs modern classics. (19)
A1B Has Jonathan Knight mentioned Dan Shields? (Yes)

- More props

We may have to wait for more props. (8)
Jason Russell has written to Robert Hutton who works for the set design team that needs an experienced carpenter. (19)
B1A Has Robert Hutton written to Jason Russell? (No)

- Potential candidates

We are replying to potential candidates. (6)
Paul Blake has answered Julian Edwards who used to work for Vision 2000 that usually employs skilled technicians. (18)
A1B Has Paul Blake answered Julian Edwards? (Yes)

- A perfect scene

The scene at the club was perfect. (7)
Jill Maxwell has thanked Kate Nelson who directs the extras casting department that recruited the group of teenagers. (18)
C1A Has the extras casting department thanked Jill Maxwell? (No)

- Special effects

We need an expert for the special effects. (8)
Tom Sutton has recommended Adam Grant who is currently working for Blue Ltd that specialises in post-production editing techniques. (19)
B2C Is Adam Grant currently working for Blue Ltd? (Yes)

- Media person

We may have found an experienced media person. (9)
Linda Berry has impressed Fiona Smith who works in the PR department that generated the newspaper coverage. (17)
B1C Has Fiona Smith impressed the PR department? (No)

D.3.3 Film Scenario - Filler Messages

- A familiar face

We could use a familiar face. (6)
Keith Smith represents Colin May who is starring in a TV series that is currently popular. (16)
A3D Is Keith Smith currently popular? (No)

- Further resignations

We can't afford further resignations. (6)
Jonathan Moore confronted Chris Philip who quit the production that ran out of money. (14)
B3D Did Chris Philip run out of money? (No)

- Being careful

We should be more careful in the future. (9)
Victoria Taylor reproached Marie Anderson who found the actress that quit the role. (13)
C3D Did the actress quit the role? (Yes)

D.4 Law Firm Scenario – Multimodal Presentation

D.4.1 Law Firm Scenario - 2 Clauses

- The new recruit

Don't forget the new recruit. (5)
Nigel Bentley called Alan Saunders who has recently graduated from London Guildhall University. (13)
B2C Has Alan Saunders recently graduated from London Guildhall University? (Yes)

- Tension

Your employees do not get along. (6)
Sally Talbot has criticised Rachel Neville who works in the contracts department. (12)
B1C Has Rachel Neville criticised the contracts department? (No)

- The new candidate

The new candidate is highly promising. (6)
Colin Miller has examined Ian Bishop who applied for the defence lawyer position. (13)
A1B Has Colin Miller examined Ian Bishop? (Yes)

- The deadline

We will not meet the deadline. (6)
Lisa Hewitt has blamed Jo Bennett who delayed the paperwork for the discrimination case. (14)
A2C Did Lisa Hewitt delay the paperwork for the discrimination case (No)

- The prosecution witness

The prosecution witness is now available. (6)
Jeremy Carlson has contacted Pete Duncan who has just returned from abroad. (12)
A1B Has Jeremy Carlson has contacted Pete Duncan? (Yes)

- The trial

We have a confirmed date for the trial. (8)
Dawn Silver has replied to Tracy Wood who works in the District Court. (13)
C1A Has the District Court replied to Dawn Silver? (No)

- Senior barristers

It is difficult to find senior barristers. (7)
Sharon Wright has faxed Emma Douglas who manages the legal department at Lloyds. (13)
B2C Does Emma Douglas manage the legal department at Lloyds? (Yes)

- Being there

You don't have to be there. (6)
Frank Ryder has emailed Andy Cooper who will conduct the hearing transcription. (12)
B1A Has Andy Cooper emailed Frank Ryder? (No)

D.4.2 Law Firm Scenario - 3 Clauses

- Witness statements

We're getting witness statements for the unfair dismissal case. (9)
Natalie Mills has interviewed Lauren Hamilton who is currently working for the internet finance company that fired the client. (19)
B2A Is Lauren Hamilton currently working for Natalie Mills? (No)

- Evidence

We should see the prosecution's evidence (6)
Kim Mortimer has approached Michelle Ellis who runs the legal department that is currently working on the murder trial. (19)
A1B Has Kim Mortimer approached Michelle Ellis? (Yes)

- International experience

We need someone with international experience (6)
Ben Austin has mentioned Nick Ellis who used to manage Security Associates that dealt with South American companies. (18)
B1C Has Nick Ellis mentioned Security Associates? (No)

- The claim

The claim will take some time. (6)
Kelly Martin has written to Lucy Daniels who knows the insurance company that is handling the bankruptcy case. (18)
A1B Has Kelly Martin written to Lucy Daniels? (Yes)

- The asylum cases

We don't need any more help with the asylum cases. (10)
Matt Oliver has answered Graham Allen who used to supervise Legal Direct that mainly focuses on immigration law. (18)
A1C Has Matt Oliver answered Legal Direct? (No)

- The progress

Everyone is pleased with the progress. (6)
Tamara Harvey has thanked Amy Louis who found the employee who remembers the details of that day. (17)
B2C Did Amy Louis find the employee? (Yes)

- Divorce lawyers

We are still short of divorce lawyers. (7)
Ray Jones has recommended Dean Morrison who is currently working for Burdon Inc. that specialises in marital disputes. (18)
C1B Has Burdon Inc. recommended Dean Morrison? (No)

- The Greenpeace case

The Greenpeace case needs a new face. (7)
Jim Porter has impressed Steve Hunt who works in the litigation department that specialises in environmental issues. (17)
B2C Does Steve Hunt work in the litigation department (Yes)

D.4.3 Law Firm Scenario - Filler Messages

- The insurance agent

You need to meet the insurance agent. (8)
Kevin Jones represents Lionel Morton who owns the car that crashed. (11)
B3D Did Lionel Morton crash? (No)

- Criticism

The judge criticised the paperwork. (5)
Christopher Bourne confronted Clive Prentice who represents the company that began the proceedings. (13)
C3D Did the company begin the proceedings? (Yes)

- Time keeping

Time keeping has got to improve. (6)
Veronica Tyler reproached Margaret Allison who missed the proceeding that ruined the case. (13)
A3D Did Veronica Tyler ruin the case? (No)

Appendix E

Preference questionnaire of the applied study

- Are you an email user?
- How long have you been using emails?

- In the experiment, did you find the visual-text easy to read (font, size)?
- Did you have enough time to read each message?
- Was the quality of sound satisfactory?
- Did you find it easy to follow the speech, given its rate?

In the combined mode of presentation:

- Did you try to ignore the speech?
- Did you try to ignore the visual-text or use it only as a back up?
- Did you consciously slow down your reading pace to synchronise it with the speech rate?
- Did you find that you needed to re-read parts of sentences?
- Did you use any other conscious strategy?

- Which mode of presentation did you find the easiest and which was the most difficult? Why?
- In the experiment, some messages were shorter than others. Which presentation supported them best and why?
- Which presentation supported long messages best and why?

- Any other comments?

- Imagining yourself using a mobile device at work, are there situations or tasks in which you could foresee speech output and combined text and speech being useful?

Appendix F

Preference data of the applied study

Index: HS - High Span, LS - Low span, Y - yes, N - no, V - visual, S - speech, MM - multimodal, BU - backup, Sync - synchronisation

Subject	Span	Email user	Usage time (Years)	Text quality	Text rate	Sound quality	Speech rate
HS_1	4.0	y	4.5	y	y	y	y
HS_2	3.5	y	4.5	y	y	y	y (for short)
HS_3	4.0	y	2.5	y	y	y (monotone)	y
HS_4	5.5	y	7.0	y	y	y	y
HS_5	4.5	y	5.0	y	y	y	y
HS_6	2.0	y	3.0	y	y	y	y
HS_7	4.5	y	9.0	y	y	y	y
HS_8	5.0	y	8.0	y	y	y (monotone)	y
HS_9	4.0	y	5.0	y	y	y	y
HS_{10}	3.5	y	10.0	y	y	y	y
HS_{11}	4.0	y	11.0	y	y	y	y
HS_{12}	3.5	y	3.5	y	y	y	y
HS_{13}	5.0	y	10.0	y	y	y	y
HS_{14}	4.5	y	2.5	y	y	y	y
HS_{15}	5.0	y	4.0	y	y	y (monotone)	y (too slow)
HS_{16}	4.0	y	10.0	y	y	y	y
Mean HS	4.2	100%	6.2	100%	100%	100%	100%
LS_{17}	2.5	y	4.5	y	y	y	y
LS_{18}	3.0	y	4.0	y	y	y	y
LS_{19}	3.0	y	10.0	y	y	y	y
LS_{20}	2.0	beginner	0.5	y	y	y	y
LS_{21}	3.0	n	0.0	y	y	y	y
LS_{22}	3.0	y	4.5	y	y	y	y
LS_{23}	2.0	n	0.0	y	y	y	y
LS_{24}	3.0	y	2.0	y	y	y	y
LS_{25}	3.0	y	1.0	y	y	y	y
LS_{26}	3.0	y	5.0	y	y	y	y
LS_{27}	3.0	y	2.0	y	y	y	y
LS_{28}	3.0	y	2.0	y	y	y	y
LS_{29}	3.0	n	0.0	y	y	y	y
LS_{30}	3.0	y	2.0	y	y	y	y
LS_{31}	2.5	y	10.0	y	y	y (monotone)	y
LS_{32}	3.0	y	5.0	y	y	y	y
Mean LS	2.8	81%	3.3	100%	100%	100%	100%

Subject	Ignored speech	Ignored text	Conscious synchronisation	2nd reading
HS_1	n	n	y	For names only
HS_2	n	n	y	not in mm, yes in visual
HS_3	n	n	y	not in mm, yes in visual
HS_4	n	y	y (but speech attend)	not in mm, yes in visual
HS_5	n	used as a BU	y (but speech attend)	not in mm, yes in visual
HS_6	n	n	y	n
HS_7	n	n	y	n
HS_8	n (but VA)	n	n	y
HS_9	n	n	y	n but did so because had the time
HS_{10}	n	n	y	probably yes
HS_{11}	at the beginning	later on, used as a BU	n	y
HS_{12}	y	n	n	n
HS_{13}	y	n	not sure	y
HS_{14}	y a bit	n	n	y
HS_{15}	y (from the middle of the sentence)	n	y (until the middle of the sentence)	names only
HS_{16}	y	n	n	y
Sum HS	y=5, n=10, ?=1	y=1, n=13, BU=1, ?=1	y=10, n=5, ?=1	mm=8, visual=4
LS_{17}	n	used as a BU	y	n
LS_{18}	n	n	y	y
LS_{19}	n	y	n	n
LS_{20}	n	n	y	n
LS_{21}	n	n	y (but found it hard)	only for long sentences in the visual condition
LS_{22}	towards the end	at the beginning	y	y for longer sentences in both visual & multimodal
LS_{23}	n	n	y (but more speech attend)	y
LS_{24}	n	used as a BU	y	not in mm, yes in visual
LS_{25}	n	n	y	names only
LS_{26}	n	n	y	y
LS_{27}	n	n	y	n
LS_{28}	n	used as a BU	y	y
LS_{29}	n	n	y	y
LS_{30}	n but VA	n	n	y
LS_{31}	y	n	y	key words only
LS_{32}	y	n	n	y
Sum LS	y=2, n=13, ?=1	y=1, n=11, BU=3, ?=1	y=13, n=3, ?=0	mm=10, visual=3
Total sum	y=7, n=23, ?=2	y=2, n+BU=28, ?=2	y=23, n=8, ?=1	mm=18, visual=7

Subject	Multimodal strategy	Other strategy
HS_1	Attended both, sync, visual channel as a BU	2nd name
HS_2	sync	rehearsal
HS_3	sync	first letter of each name and verbs
HS_4	visual text interfered with listening	2nd name plus imagery
HS_5	sync but was listening more to the speech	names and position
HS_6	sync but was attentive more to the visual channel	names and verbs
HS_7	sync	names and verbs towards the end
HS_8	visual attend. Didn't find speech interfering	2nd name
HS_9	reading quickly then sync	names and verbs
HS_{10}	sync	2nd name plus imagery (by the meaning of the name)
HS_{11}	At first was visual attend, then speech attend and visual BU	none
HS_{12}	visual attend	ignored priming sentences, used names and position
HS_{13}	visual attend	1st reading for comp, 2nd reading focused on the 2nd name
HS_{14}	visual attend	names and verbs, but mainly 2nd name
HS_{15}	Mixed attention: listening to the names, reading ahead, visual BU	first letter of each surname and order in which they appeared
HS_{16}	visual attend, tried to block the speech	names and verbs, sometimes used initials
Sum HS		Names (including 2nd name):13, 2nd name=6
LS_{17}	Switched strategies; speech attend plus visual BU	names and position
LS_{18}	used traces of speech, sync	names and verbs (not in the MM)
LS_{19}	speech attend, used traces of speech, no sync	repeating the whole sentence
LS_{20}	sync	none
LS_{21}	The MM condition did not enable to re-read the sentence	names and verbs
LS_{22}	At first was speech attend, then sync but visual attend	visualised the structure of the sentence
LS_{23}	sync but was attentive more to the speech	2nd name
LS_{24}	speech attend, used traces of speech, no sync	names and verbs
LS_{25}	sync	2nd name
LS_{26}	visual attend, reading quickly then sync, finally visual BU	2nd name
LS_{27}	sync	2nd name, ignored the last bit of the sentence
LS_{28}	sync, speech attend, visual backup	ignored priming sentences, used names and verbs
LS_{29}	visual attend, reading quickly then sync, finally visual BU	names and order, ignored surnames
LS_{30}	visual attend, picking key words in the speech channel	1st reading for comp, 2nd for names and verbs
LS_{31}	visual attend, reading quickly then sync, finally visual BU	echoic memory plus names & verbs
LS_{32}	visual attend, reading quickly then sync, finally visual BU	Ignored titles and priming sentences
Sum LS		Names (including 2nd name):12, 2nd name=4
Total sum		Names (including 2nd name):25, 2nd name=10

Subject	Easy condition	Difficult condition	Medium condition	Easy for short	Easy for long
HS_1	mm	speech	visual	mm	mm
HS_2	mm	speech	visual	mm	mm
HS_3	visual	speech	mm	speech	visual
HS_4	speech	mm	visual	visual	speech
HS_5	speech	visual	mm	mm	speech
HS_6	visual	speech	mm	visual	mm
HS_7	visual	speech	mm	speech	visual
HS_8	visual	speech	mm	same	visual
HS_9	mm	speech	visual	mm	mm
HS_{10}	mm	speech	visual	mm	?
HS_{11}	speech	visual	mm	visual	?
HS_{12}	mm	visual	speech	same	mm
HS_{13}	speech	mm	visual	speech	speech
HS_{14}	mm	visual	speech	?	mm
HS_{15}	mm	visual	speech	same	mm
HS_{16}	visual	mm	speech	?	?
Sum HS	mm=7, v=5, s=4	mm=3, v=5, s=8	mm=6, v=6, s=4	mm=5, v=3, s=3, same=3, ?=2	mm=7, v=3, s=3, same=0, ?=3
LS_{17}	mm	visual	speech	visual	mm
LS_{18}	visual	speech	mm	visual	mm
LS_{19}	speech	mm	visual	visual	speech
LS_{20}	mm	visual	speech	mm	mm
LS_{21}	visual	speech	mm	mm	visual
LS_{22}	visual	mm	speech	visual	?
LS_{23}	mm	visual	speech	mm	mm
LS_{24}	visual	mm	speech	same	visual
LS_{25}	visual	speech	mm	visual	visual
LS_{26}	mm	speech	visual	?	?
LS_{27}	mm	speech	visual	speech	mm
LS_{28}	mm	visual	speech	visual	mm
LS_{29}	visual	mm	speech	visual	visual
LS_{30}	visual	speech	mm	speech	visual
LS_{31}	mm	speech	visual	speech	mm
LS_{32}	mm	speech	visual	same	mm
Sum LS	mm=8, v=7, s=1	mm=4, v=4, s=8	mm=4, v=5, s=7	mm=3, v=7, s=3, same=2, ?=1	mm=8, v=5, s=1, same=0, ?=2
Total sum	mm=15, v=12, s=5	mm=7, v=9, s=16	mm=10, v=11, s=11	mm=8, v=10, s=6, same=5, ?=3	mm=15, v=8, s=4, same=0, ?=5

Subject	Other comments	Usefulness
HS_1	none	no need in PDA devices
HS_2	none	Flexibility, convenient to listen, useful for long messages
HS_3	the user should be able to control presentation rate	while driving, not in the office
HS_4	Trained to pick up messages on the mobile. Good listener	
HS_5	none	when being mobile. S rate is good enough to understand long messages.
HS_6	wouldn't like to read a MM book	home use, not office use
HS_7	none	wouldn't use it, intrusive
HS_8	the audio was good	Provides flexibility. Visual BU
HS_9	a learning effect	flexibility, speech for driving, would prefer visual only
HS_{10}	none	printing or ringing telephone numbers while listening to the answering machine
HS_{11}	none	wouldn't use it, will stick to the text
HS_{12}	none	free ones hands, visual BU, printing
HS_{13}	problems with the first message	on a train
HS_{14}	none	when being mobile. When hands are busy, poor eyesight, poor lighting conditions
HS_{15}	none	when busy or tired, listening is easier. Visual serves as a BU
HS_{16}	none	wouldn't use it, those devices introduce further stress
Sum HS	None: 10	Not useful: 4
LS_{17}	none	not useful
LS_{18}	none	flexibility, check-up mechanism, creates background noise
LS_{19}	none	not useful, annoying
LS_{20}	was very nervous at the beginning	
LS_{21}	It all depends on reading speed - she's a fast reader	when information needs to be very precise (e.g., telephone numbers, names)
LS_{22}	none	when doing something else (answering machine). Wouldn't use both modalities
LS_{23}	none	an addition to the mobile phone
LS_{24}	none	wouldn't use it, will stick to the text
LS_{25}	none	wouldn't use it, will stick to the text
LS_{26}	none	visual output in a noisy environment, flexibility in multi-tasking
LS_{27}	none	flexibility in busy environment
LS_{28}	none	no
LS_{29}	none	no
LS_{30}	none	would prefer visual only
LS_{31}	in real life, you know one person	disability, when one is away from the desk, spoken messages save time in a group-work
LS_{32}	none	voice mail attracts attention, MM frees you up
Sum LS	None: 13	Not useful: 8
Total sum	None: 23	Not useful: 12